Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

# Cost efficiency of repeated cluster surveys


by Stanislav Kolenikov and Gustavo Angeles

June 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 75-94
Statistics Canada, Catalogue No. 12-001-X



# Cost efficiency of repeated cluster surveys

**Stanislav Kolenikov and Gustavo Angeles** [1]

Abstract

We analyze the statistical and economic efficiency of different designs of cluster surveys collected in two consecutive time periods, or waves. In an independent design, two cluster samples in two waves are taken independently from one another. In a cluster-panel design, the same clusters are used in both waves, but samples within clusters are taken independently in two time periods. In an observation-panel design, both clusters and observations are retained from one wave of data collection to another. By assuming a simple population structure, we derive design variances and costs of the surveys conducted according to these designs. We first consider a situation in which the interest lies in estimation of the change in the population mean between two time periods, and derive the optimal sample allocations for the three designs of interest. We then propose the utility maximization framework borrowed from microeconomics to illustrate a possible approach to the choice of the design that strives to optimize several variances simultaneously. Incorporating the contemporaneous means and their variances tends to shift the preferences from observation-panel towards simpler panel-cluster and independent designs if the panel mode of data collection is too expensive. We present numeric illustrations demonstrating how a survey designer may want to choose the efficient design given the population parameters and data collection cost.

Key Words: Longitudinal study; Cluster samples; DHS; NHIS.

## 1. Introduction

To analyze the dynamics of social, behavioral or population health phenomena, researchers and policymakers need to obtain information on characteristics of the population on multiple occasions. Complex design surveys are the most frequently used sources of information for large populations, such as a country as a whole. Besides the standard considerations in single-shot surveys, *e.g.*, stratification and clustering, other issues may be important in surveys collected over two or more time periods. In such surveys, the total cost and the total survey error are affected by an overlap among consecutive samples, (informative) sample attrition, time-in-sample or conditioning effects, and other dynamic factors.

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 achieved by administering the survey to the same sampling and/or observation units. In longitudinal surveys, the same observation units (individuals, households) are revisited for several periods, potentially indefinitely many periods (the US Panel Study of Income Dynamics (PSID), British Household Panel Study (BHPS) and others). A compendium of information on the longitudinal studies can be found at the Institute for Social and Economics Research web site, http://iser.essex.ac.uk/ulsc/keeptrack/index.php). In rotating panel surveys, the observation units are recruited into the sample for a few periods, then rotated out of the sample, and surveyed again at a later time. Examples of rotating panel surveys include the US Current Population Survey (CPS) (Binder and Hidiroglou 1988, Eckler 1955, Rao and Graham 1964) and a number of environmental surveys (Fuller 1999, McDonald 2003, Scott 1998). Yet another option is to use the same primary sampling units (PSUs) in different waves, but sample the observation units (secondary sampling units, SSUs) independently. Surveys collected in this way include international Demographic and Health Surveys (DHS) and the US National Health Interview Survey (NHIS).

We shall concentrate on surveys collected in two time periods, or waves, using a two-stage cluster design in each wave of data collection. We consider three possible designs differing in the amount and depth of overlap of sampling units over time. The sample designer can simply ignore any possible effects arising from the sample overlap, and take two independent samples in two periods of time. We shall refer to this design as the *independent* design. Alternatively, the sample designer may find it beneficial to recycle the PSUs from one wave to another. If the designer finds it difficult to track the SSUs from one wave to another, the subsamples within clusters can be taken independently in two waves of data collection. We shall refer to this design as the *cluster-panel* design. If an utmost precision is essential, the fully longitudinal design will attempt to locate all individuals who responded in the first wave, and solicit the second interview. To distinguish this design from the cluster-panel design, we shall refer to it as the *observation-panel* design.

1. Stanislav Kolenikov, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100, U.S.A. E-mail: kolenikovs@missouri.edu; Gustavo Angeles, Associate Director of the Center for Evaluation Research, National Institute of Public Health, Mexico, Mexico. E-mail: gangeles@insp.mx.

A particular aspect that we found important in survey management, but underaddressed in the existing literature, is the implementation cost (Groves 1989). The traditional cost models such as those used in derivation of Neyman-Tchuprow optimal allocation design (Neyman 1938) can be extended to include terms related to the cost of the first visit to the cluster and ultimate observation unit, as well as the cost of consecutive visits. The cost of revisiting the cluster is likely to be lower on the second occasion. There is no need to create new maps and set up frames. The same interviewers can be used to conduct interviews in subsequent waves of data collection. Cooperation with community leaders has been established earlier, if it is important, as it is in some traditional societies. The effect of the panel mode of data collection at the individual level is less clear. If the household that was interviewed in earlier waves moved out and would have to be located, possibly in different geographic area, the (average) cost of the panel interview goes up. The likelihood of such circumstances increases with longer intervals between surveys typical for the developing countries surveys: the intervals between waves of DHS are usually about 5-7 years. On the other hand, if a less expensive interview mode can be used after the first round, (*e.g.*, a phone interview instead of the personal visit), the cost of the panel interview goes down.

This paper brings together statistical and economic considerations in the choice of the appropriate design and its parameters. We assume the survey designer can be interested in estimating the change in the population mean between two time periods, and/or the means themselves. We introduce a sketchy population in Section 2, and compute the design variances of the means and their differences for the three sampling designs of our interest.

To incorporate economic aspects of data collection, we introduce a relatively simple cost model for a repeated cluster survey in Section 3. We set up and solve optimization problems to obtain the optimal sample sizes for the three considered designs. By plugging in the estimates of the statistical parameters (variances and autocorrelations) and cost components (cluster-level and individual-level costs), the survey designer can compare the numeric values of the variances to choose the best design. Section 4 illustrates this approach and shows that each of the designs may be the best one, depending on the parameter values. The intuitive results (*e.g.*, the higher cost of data collection and lower autocorrelations of the observed characteristics make panel modes of data collection less appealing) are given an analytic justification and quantitative backing.

While Sections 2-4 deal with the efficiency in estimating the difference in means only, more realistic goals of data collection efforts would include contemporaneous characteristics and their variances. To this end, Section 5 introduces a utility maximization framework describing the survey designer's choice of the sampling scheme. This framework provides an aggregated objective function that combines several design criteria. The results are again as expected: if the more expensive panel modes of data collection result in smaller sample sizes, the estimates of the means are less efficient than in simpler designs. The only way to justify these efficiency losses is by a drastic improvement in the estimation of the difference that can only occur with higher autocorrelations. Such effects are also illustrated in Section 5. Section 7 concludes. Proofs are given in the Appendix.

## 2. Design variances

Let the population consist of $N$ clusters, or PSUs, in both time periods, and each cluster consist of $M$ individuals, or SSUs. Out of these, an SRS of $1 < n_t \le N$ clusters is taken at time $t = 1, 2,$ and an SRS of $1 < m_t \le M$ individuals is taken in each cluster that is present in the sample at time $t$. Let the index $i$ denote PSUs, and the index $j$, SSUs. Thus the typical measurement will be denoted as $Y_{tij}$ in the population, and $y_{tij}$ in the sample. The population totals $T[\cdot]$ and their estimates $t[\cdot]$ can then be found as follows:

cluster total:

$$T_{ti\cdot}[Y] = \sum_{j=1}^{M} Y_{tij},\ t_{ti\cdot}[y] = \frac{M}{m}\sum_{j=1}^{M} y_{tij},$$

population total:

$$T_{t\cdot\cdot}[Y] = \sum_{i=1}^{N} Y_{ti\cdot},\ t_{t\cdot\cdot}[y] = \frac{N}{n}\sum_{i=1}^{N} t_{ti\cdot}[y]. \qquad (2.1)$$

The means per observation units are

$$\bar{Y}_{ti\cdot} = \frac{1}{M}\sum_{i=1}^{M} Y_{tij} = \frac{T_{ti\cdot}[Y]}{T_{ti\cdot}[1]},\ \bar{y}_{ti\cdot} = \frac{1}{m}\sum_{i=1}^{m} y_{tij} = \frac{t_{ti\cdot}[y]}{t_{ti\cdot}[1]},$$

$$\bar{Y}_{t\cdot\cdot} = \frac{T_{t\cdot\cdot}[Y]}{T_{t\cdot\cdot}[1]} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{tij}}{NM},\ \bar{y}_{t\cdot\cdot} = \frac{t_{t\cdot\cdot}[y]}{t_{t\cdot\cdot}[1]} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{tij}}{nm}. \qquad (2.2)$$

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

$$S_t^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{tij} - \bar{Y}_{t\cdot\cdot})^2}{N\bar{M} - 1}, \qquad (2.3)$$

$$S_{twi}^2 = \frac{\sum_{j=1}^{M}(Y_{tij} - \bar{Y}_{ti\cdot})^2}{M - 1},\ \bar{S}_{tw}^2 = \frac{1}{N}\sum_{i=1}^{N} S_{twi}^2, \qquad (2.4)$$

$$S_{tb}^2 = \frac{\sum_{i=1}^{N}(\overline{Y}_{ti\cdot} - \overline{Y}_{t\cdot\cdot})^2}{N-1}. \quad (2.5)$$

The characteristic of primary interest is the change in the means,

$$D = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{1\cdot\cdot}, \quad (2.6)$$

estimated by

$$d = \overline{y}_{2\cdot\cdot} - \overline{y}_{1\cdot\cdot}. \quad (2.7)$$

An attractive property of this estimator for analysts and data users is its internal consistency: the estimator of the difference is the difference of the estimators. If the samples in consecutive periods overlap only partially, then composite or GLS estimators (Fuller 1999, Hansen, Hurwitz and Madow 1953, Patterson 1950, Rao and Graham 1964, Wolter 2007) have better efficiency.

In what follows, we assume all sampling procedures to be simple random sampling without replacement. For the contemporaneous mean, the variance is given by (Cochran 1977, Th. 10.1):

$$\mathrm{V}[\overline{y}_{t\cdot\cdot}] = \left(1 - \frac{n}{N}\right)\frac{S_{tb}^2}{n} + \left(1 - \frac{m}{M}\right)\frac{\overline{S}_{tw}^2}{nm}. \quad (2.8)$$

For simplicity and clarity of exposition, we shall often be making an assumption of symmetric conditions:

$$S_{1wi}^2 = S_{2wi}^2 = S_{wi}^2,\ \overline{S}_{1w}^2 = \overline{S}_{2w}^2 = \overline{S}_w^2,\ S_{1b}^2 = S_{2b}^2 = S_b^2. \quad (2.9)$$

Analytic derivations are possible without these assumptions, but become extremely cumbersome. Besides, it is unrealistic to think that the survey designer could know the characteristics of the future population. Thus (2.9) should be viewed as a reasonable working model.

### 2.1 Independent design

*Proposition* 1. *Let $n_1$ out of $N$ clusters and $m_1$ out of $M$ observation units in selected clusters be taken without replacement at time $t = 1$. Let $n_2$ out of $N$ clusters and $m_2$ out of $M$ observation units in selected clusters be taken without replacement at time $t = 2$, with sampling performed independently from that at time $t = 1$. Then*

$$\mathrm{V}_\iota(d) = \left(1 - \frac{n_1}{N}\right)\frac{S_{1b}^2}{n_1} + \left(1 - \frac{n_2}{N}\right)\frac{S_{2b}^2}{n_2} + \left(1 - \frac{m_1}{M}\right)\frac{\overline{S}_{1w}^2}{n_1 m_1} + \left(1 - \frac{m_2}{M}\right)\frac{\overline{S}_{2w}^2}{n_2 m_2}. \quad (2.10)$$

The result follows immediately from (2.8) by independence of the two samples. The subindex of the variance $\iota$ stands for the "independent design". Under the symmetric conditions of (2.9), if the sample sizes are the same in two periods, $n_1 = n_2 = n$ and $m_1 = m_2 = m$, then

$$\mathrm{V}_{e,\iota}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\overline{S}_w^2}{nm}, \quad (2.11)$$

where the subindex $e, \iota$ stands for "equal variances, independent design".

### 2.2 Cluster-panel design

*Proposition* 2. *Let $n$ out of $N$ clusters be sampled without replacement in the first period and be used in both time periods. Let $m$ out of $M$ observation units be sampled without replacement independently in two periods. Then*

$$\mathrm{V}_c[d] = \left(1 - \frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right)\frac{\overline{S}_{1w}^2 + \overline{S}_{2w}^2}{nm},$$

$$\rho^{\mathrm{I}} = \frac{1}{S_{1b} S_{2b}(N-1)}\sum_{i=1}^{N}(\overline{Y}_{1i\cdot} - \overline{Y}_{1\cdot\cdot})(\overline{Y}_{2i\cdot} - \overline{Y}_{2\cdot\cdot}). \quad (2.12)$$

Here, subindex $c$ stands for the "cluster-panel design", and $\rho^{\mathrm{I}}$ is the intertemporal correlation, or autocorrelation, of the cluster means. The superscript I denotes the first stage of sampling. If $\rho^{\mathrm{I}}$ is positive, then the cluster-panel design is more efficient than the independent design for fixed values of $n$ and $m$. Under the symmetry conditions,

$$\mathrm{V}_{e,c}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2(1-\rho^{\mathrm{I}})}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\overline{S}_w^2}{mn}, \quad (2.13)$$

where the subindex $e, c$ stands for the "equal variances, cluster-panel design".

### 2.3 Observation-panel design

*Proposition* 3. *Let $n$ out of $N$ clusters and $m$ out of $M$ observation units be sampled without replacement in the first period and be used in both time periods. Then*

$$\mathrm{V}_o[d] = \left(1 - \frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right)\frac{\overline{S}_{1w}^2 + \overline{S}_{2w}^2 - 2\rho^{\mathrm{II}}\overline{S}_{1w}\overline{S}_{2w}}{nm},$$

$$\rho^{\mathrm{II}} = \frac{1}{\overline{S}_{1w}\overline{S}_{2w}N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{1ij} - \overline{Y}_{1i\cdot})(Y_{2ij} - \overline{Y}_{2i\cdot}). \quad (2.14)$$

Subindex $o$ stands for the "observation-panel design". Under the assumption of symmetric conditions,

$$V_{e,o}[d] = 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{nm},$$

$$\rho^{II} = \frac{1}{\bar{S}_w^2 N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{1ij}-\bar{Y}_{1i\cdot})(Y_{2ij}-\bar{Y}_{2i\cdot}) \quad (2.15)$$

with corresponding $e,o$ subindex for the "equal variances, observation-panel design".

Here, $\rho^{II}$ is the intertemporal correlation, or autocorrelation, of the individual observations within clusters. The superscript II stands for the second stage of sampling. If $\rho^{II}$ is positive, then the observation-panel design is more efficient than the cluster-panel design for fixed values of $n$ and $m$.

How are the two autocorrelations that appear in (2.15) related? Conceptually, one can think of any number of possible relations between them. Let us introduce a superpopulation model

$$Y_{tij} = \mu_t + a_{ti} + \varepsilon_{tij}, \quad E_\xi[a_{ti}] = 0, \quad E_\xi[\varepsilon_{tij}] = 0, \quad (2.16)$$

in which $a_{ti}$ and $\varepsilon_{sij}$ are independent of one another for all $s, t = 1, 2$. The subindex $\xi$ stands for the superpopulation model expectations. The case of $\rho^{I} = 0$ and $\rho^{II} = 1$ occurs when the changes in the cluster means occur independently between clusters ($E_\xi[a_{1i}a_{2i}] = 0$), but the individuals retain their positions within the cluster, $\varepsilon_{1ij} = \varepsilon_{2ij}$. The case of $\rho^{I} = 1$ and $\rho^{II} = 0$ occurs when the cluster random effects are the same in both periods, $a_{1i} = a_{2i}$, while the individual random effects are uncorrelated ($E_\xi[\varepsilon_{1ij}\varepsilon_{2ij}] = 0$). Neither of these situations is entirely realistic. However, it can probably be expected that the individual, rather than the cluster, dynamics are a more important source of variation over time, thus making the relations $\rho^{II} \geq \rho^{I} \geq 0$ the most plausible ones. We shall study in numeric examples of Sections 4 and 5 the extent to which the choice of the best design is sensitive to the relation between the two correlations.

## 3. Costs for repeated cluster samples

In this section we shall analyze the cost efficiency of cluster 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 Kish (1995, Section 8.3B), Thompson (1992, Section 12.5), and Lehtonen and Pahkinen (2004). More mathematical details are available in Hansen *et al.* (1953, volume II, Section 6.11), with the variance formulas corrected for finite populations.

### 3.1 Notation and cost models

Let us assume the following cost structure, which is an extension of Kish (1995) for repeated surveys:

- $c_1^{I}$ is the cluster level cost at time $t = 1$ for clusters that are used *in the first wave only*;
- $c_2^{I}$ is the cluster level cost for a *new* cluster at time $t = 2$;
- $c_{12}^{I}$ is the cluster level cost for clusters in which the data are collected in both periods $t = 1$ and $t = 2$ (PSU panel cost);
- $c_1^{II}$ is the individual level cost at time $t = 1$ for individuals that are observed *in the first wave only*;
- $c_2^{II}$ is the individual level cost at time $t = 2$ for individuals that are observed *in the second wave only*;
- $c_{12}^{II}$ is the individual level cost if the unit is observed in both periods in the observation-panel design (SSU panel cost);
- $C_0$ is the total budget allocated to the field work in both time periods.

Roman superscripts denote the sampling stage. Arabic subscripts correspond to the occasion at which the sample is taken. The cluster level costs include the cost of sampling the clusters, obtaining the PSU maps, collecting community data, local interviewer training, *etc.* The individual level costs are mostly those of the personal interviews with the ultimate observation units. The total cost $C_0$ is thought of as the variable cost of the survey that is directly related to the number of sampled units. Fixed cost, such as the cost of preparing the survey instrument and other organization-level costs are not part of $C_0$.

### 3.2 Independent design

The budget constraint for the independent design is given by

$$C_0 = c_1^{I} n_1 + c_1^{II} n_1 m_1 + c_2^{I} n_2 + c_2^{II} n_2 m_2. \quad (3.1)$$

The first two terms are the costs of the first wave of data collection, and the last two terms, of the second wave.

*Proposition* 4. *If the survey setting parameters are the same in the two time periods*:

$$c_1^{I} = c_2^{I} = c^{I}, \; c_1^{II} = c_2^{II} = c^{II}, \quad (3.2)$$

*then the optimal sample sizes and the resulting variances are given by*

$$m = \sqrt{\frac{c^{\mathrm{I}}}{c^{\mathrm{II}}}\frac{\overline{S}_w^2}{S_b^2 - \overline{S}_w^2/M}},$$

$$n = \frac{C_0}{2\{c^{\mathrm{I}} + [c^{\mathrm{I}}c^{\mathrm{II}}\overline{S}_w^2/(S_b^2 - \overline{S}_w^2/M)]^{1/2}\}},$$

$$\mathrm{V}_{e,\iota}[d] = \frac{4\left[c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}}c^{\mathrm{II}}\overline{S}_w^2/(S_b^2 - \overline{S}_w^2/M)}\right]}{C_0} \times \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}\frac{S_b^2 - \overline{S}_w^2/M}{\overline{S}_w^2}} - \frac{1}{M}\right)\overline{S}_w^2\right] - \frac{2}{N}S_b^2. \quad (3.3)$$

In equations (3.3), the sample sizes $n$ and $m$ are treated as continuous variables. In practice, the nearest integer should be used, with a minimum of 2 necessary to estimate the appropriate variance component, and the maxima of $N$ and $M$, respectively.

The number of observations sampled within a cluster depends only on the relative costs at the cluster and the observation level, $c^{\mathrm{I}}/c^{\mathrm{II}}$, and relative variances $S_b^2/\overline{S}_w^2$, or equivalently the intraclass correlation. Greater interview cost $c^{\mathrm{II}}$ prevents the sample designer from using more observations: an increase in $c^{\mathrm{II}}$ leads to a decrease in both $m$ and $n$. Greater cluster-level cost leads to redistribution of the sampled units: $n$ decreases with $c^{\mathrm{I}}$, while $m$ increases with it. Greater within-cluster variance $\overline{S}_w^2$ necessitates a greater number of observations $m$ to be taken within a cluster to maintain overall precision. Greater between-cluster variance $S_b^2$ necessitates a greater number of clusters $n$ to be sampled. Finally, the total survey budget $C_0$ affects the number of clusters $n$, but not the subsample size $m$. As a result, the variance of $d$ is inversely proportional to $C_0$.

The non-symmetric situation can be treated as a by-product of the first order conditions derived in the proof (see Appendix). However, no analytic solution is available in that case.

### 3.3 Cluster-panel design

The budget constraint for the cluster-panel design is given by

$$C_0 = c_{12}^{\mathrm{I}}n + c_1^{\mathrm{II}}nm_1 + c_2^{\mathrm{II}}nm_2. \quad (3.4)$$

The first term is the cluster-level cost associated with the sample design, and the remaining two terms are the costs of collecting individual-level data in the first and the second waves, respectively.

*Proposition* 5. *The sample sizes for the cluster-panel design are given by*

$$m_1 = 2C_0\Big/c_1^{\mathrm{II}}\left(1 + \frac{\overline{S}_{2w}^2}{\kappa\overline{S}_{1w}^2} + \sqrt{D}\right),$$

$$m_2 = \kappa m_1,$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}}m_1 + c_2^{\mathrm{II}}m_2},$$

$$\kappa = \sqrt{\frac{c_1^{\mathrm{II}}}{c_2^{\mathrm{II}}}\frac{\overline{S}_{2w}^2}{\overline{S}_{1w}^2}}, \quad (3.5)$$

*provided that*

$$D = \left(1 + \frac{\overline{S}_{2w}^2}{\kappa\overline{S}_{1w}^2}\right)^2 + 8\frac{(1-\rho^{\mathrm{I}})S_b^2C_0}{\overline{S}_{1w}^2c_1^{\mathrm{II}}} - 4\frac{C_0}{c_1^{\mathrm{II}}M}\left(1 + \frac{\overline{S}_{2w}^2}{\kappa\overline{S}_{1w}^2}\right) \geq 0.$$

The variance of the difference estimator is found by plugging these expressions into (2.13). Under the assumptions of symmetric conditions in two rounds of the survey (2.9) and (3.2),

$$D = 4 - 8\frac{C_0}{Mc^{\mathrm{II}}} + 8\frac{(1-\rho^{\mathrm{I}})S_b^2C_0}{\overline{S}_w^2c^{\mathrm{II}}},$$

$$m_1 = m_2 = m = \frac{C_0}{c^{\mathrm{II}} + \sqrt{(c^{\mathrm{II}})^2 - \frac{2c^{\mathrm{II}}C_0}{M} + \frac{2(1-\rho^{\mathrm{I}})S_b^2C_0c^{\mathrm{II}}}{\overline{S}_w^2}}},$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + 2c^{\mathrm{II}}m} = \frac{C_0}{c_{12}^{\mathrm{I}} + 2C_0\Big/\left[1 + \sqrt{1 - \frac{2c^{\mathrm{II}}C_0}{Mc^{\mathrm{II}}} + \frac{2(1-\rho^{\mathrm{I}})S_b^2C_0}{\overline{S}_w^2c^{\mathrm{II}}}}\right]},$$

and $\mathrm{V}_{e,c}[d]$ can be found from (2.13).

Interestingly, the number of the SSUs depends on the SSU costs $c^{\mathrm{II}}$, but not on the PSU costs $c_{12}^{\mathrm{I}}$. An increase in the intracluster correlation, or increase in $S_b^2$, or decrease in $S_w^2$, predictably leads to decrease in the optimal number of SSUs and increase in the optimal number of PSUs. The dependence of the design parameters on the survey budget $C_0$ is non-trivial. For very small surveys, the number of units per cluster is proportional to $C_0$, and the number of clusters is not affected by $C_0$. Indeed, if the characteristic demonstrates strong correlation between time periods, it would be preferable to get accurate estimates of the cluster means, and good accuracy of the overall difference estimator will follow. To put it differently, the first term in (2.13) is relatively small by virtue of the positive correlation coefficient $\rho^{\mathrm{I}}$, and the second term is inversely proportional

to $C_0$. For large surveys, $D \propto C_0$, so both the number of units per cluster and the number of clusters are proportional to $\sqrt{C_0}$. The first term in (2.13) is then inversely proportional to $\sqrt{C_0}$, and the second term is inversely proportional to $C_0$. An increase in the budget of the survey will affect all terms, although to a different extent.

### 3.4 Observation-panel design

The budget constraint for the observation-panel design is given by

$$C_0 = c_{12}^{\mathrm{I}} n + c_{12}^{\mathrm{II}} nm. \tag{3.6}$$

The first term is the cluster-level cost, and the second term is the cost of individual interviews.

*Proposition* 6. *The optimal sample sizes for the observation-panel design are given by*

$$m = \sqrt{\frac{c_{12}^{\mathrm{I}}}{c_{12}^{\mathrm{II}}} \frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}},$$
$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + \sqrt{\dfrac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}}}. \tag{3.7}$$

*The design variance of the resulting difference estimator is*

$$\begin{aligned} \mathrm{V}_{e,o}[d] = \frac{2}{C_0}\Bigg\{ & (1-\rho^{\mathrm{I}}) S_b^2 c_{12}^{\mathrm{I}} \\ & + (1-\rho^{\mathrm{II}})\bar{S}_w^2 \sqrt{c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}} \frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} \\ & + \left[ (1-\rho^{\mathrm{I}}) S_b^2 - \frac{1}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2 \right] \\ & \times \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}} \\ & + (1-\rho^{\mathrm{II}})\bar{S}_w^2 \left( c_{12}^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M} \right) \Bigg\} - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}. \end{aligned} \tag{3.8}$$

The sample size expressions (3.7) resemble the ones for the independent design, equation (3.3), with the cost of data collection in a single wave replaced by the cost of panel data collection, and the variance components $S_b^2$ and $\bar{S}_w^2$ replaced by $(1-\rho^{\mathrm{I}})S_b^2$ and $(1-\rho^{\mathrm{II}})\bar{S}_w^2$. The second stage sampling size $m$ only depends on the relative cost at the cluster and observation levels, and on the ratio of the variance components augmented by the autocorrelations. Hence, like in the independent design, the dependency of the sample size on the scale of the survey is only through $n \propto C_0$, and the variance of the difference decreases inversely proportional to $C_0$.

Extending the relations between the functional forms of equations (3.3) and (3.8), we can establish the general relations between the two designs:

*Proposition* 7. *If* $M \gg 1$ *and* $N \gg 1$, *then* $\mathrm{V}_{e,\iota}[d] \gtrless \mathrm{V}_{e,o}[d]$ *if*

$$2\left(\sqrt{c^{\mathrm{I}} S_b^2} + \sqrt{c^{\mathrm{II}} \bar{S}_w^2}\right)^2 \gtrless \left[\sqrt{c_{12}^{\mathrm{I}}(1-\rho^{\mathrm{I}})S_b^2} + \sqrt{c_{12}^{\mathrm{II}}(1-\rho^{\mathrm{II}})\bar{S}_w^2}\right]^2. \tag{3.9}$$

Unfortunately, the variance for the cluster-panel design that can be obtained by combining the results of Proposition 5 with (2.13), does not permit an equally lucid comparison.

## 4. Numeric illustration

To illustrate how the characteristics of population (variances and autocorrelations) and the data collection process (costs) affect the choice of the most efficient design, we consider a numeric example. Let us choose the basic setup with symmetric conditions, and let the parameter values be:

$$\begin{aligned} & N = 10{,}000, \quad M = 1{,}000, \quad S_b = 100, \\ & S_w = 400, \quad \rho^{\mathrm{I}} = 0.1, \quad \rho^{\mathrm{II}} = 0.35, \\ & c_1^{\mathrm{II}} = c_2^{\mathrm{II}} = 1, \quad c_{12}^{\mathrm{II}} = 3, \quad c_1^{\mathrm{I}} = c_2^{\mathrm{I}} = 10, \\ & c_{12}^{\mathrm{I}} = 18, \quad C_0 = 20{,}000. \end{aligned} \tag{4.1}$$

The cost structure implies that the cost of collecting the initial information for a cluster is the cost of ten interviews, while the cost of the followup in the same cluster is only eight interviews. On the other hand, getting the second interview with the same unit is twice as expensive as getting the first interview.

With these parameters, the sample sizes and design variances are:

$$\begin{aligned} & m_{e,\iota} = 12, \quad m_{e,c} = 12, \quad m_{e,o} = 8, \\ & n_{e,\iota} = 455, \quad n_{e,c} = 476, \quad n_{e,o} = 476, \\ & m_{e,\iota} n_{e,\iota} = 5{,}460, \quad m_{e,c} n_{e,c} = 5{,}712, \quad m_{e,o} n_{e,o} = 3{,}808, \\ & \mathrm{V}_{e,\iota}[d] = 99.86, \quad \mathrm{V}_{e,c}[d] = 91.37, \quad \mathrm{V}_{e,o}[d] = 90.20. \end{aligned} \tag{4.2}$$

The observation-panel design is 1.2% more efficient than the cluster-panel design, and 10.7% more efficient than the independent design. However, it has a notably smaller total sample size, only $2/3$ of the cluster-panel design sample size and 70% of the independent design sample size.

Of course these finding are highly specific to the parameters of the population and the cost structure. Can we describe general patterns of how the variances, and hence the relative efficiency of different designs, change with those parameters? The variances in (4.2) are derived from 13 parameters given in (4.1), and it is difficult to make meaningful statements about all of these parameters simultaneously. Below, we shall attempt to provide two-dimensional cross-sections of this 13-dimensional space and give graphical illustrations of the variability of the design variances, and hence the domains of optimality of each design, as we vary two parameters at a time. We provide the graphs of variances of the designs involved (typically, the cluster-panel design with dotted lines, the observation-panel design with dashed lines, and the independent design with dash-dotted lines. For most plots, the independent design is not affected by the variations of the parameters that make up the axis of the plots, and hence omitted). We also show the relative efficiency of different designs, marking the domains of the parameter space in yellow/light gray if the independent design is the most efficient one; in green/medium gray if the cluster-panel design is the most efficient one; and in purple/dark gray if the observation-panel design is the most efficient one (R code used to produce graphs is available at http://web.missouri.edu/~kolenikovs/SMJ2011/).

Figure 1 shows how the design variances, and hence the most efficient design, vary with the panel costs of the PSU and SSU, $c_{12}^{\mathrm{I}}$ and $c_{12}^{\mathrm{II}}$. Obviously, these variations do not affect the variance of the independent design, which serves as a benchmark. Also, the variations in $c_{12}^{\mathrm{II}}$ do not affect the performance of the cluster-panel design, which corresponds to the dotted vertical iso-variance lines on the left panel. The dashed downward sloping lines are the iso-variance lines for the observation-panel design. Note that the lower left corner of the graph corresponds to the free lunch situation in which the second wave of data collection does not cost anything: the panel costs are equal to the single period cost, $c_{12}^{\mathrm{I}} = c_{1}^{\mathrm{I}}$, $c_{12}^{\mathrm{II}} = c_{1}^{\mathrm{II}}$. When the costs of the panel data collection are prohibitively high (the upper right corner of the graph), the independent design is the most efficient one. The point where all three designs have the same variances is $c_{12}^{\mathrm{I}} = 22$, $c_{12}^{\mathrm{II}} = 3.05$; *i.e.*, the cost of the second interview is 2.05 higher than the cost of the first interview, and the cluster-level costs in the second wave are 20% higher than in the first wave. Still, a positive autocorrelation justifies the reduction in the sample size of the observation-panel design as compared to the independent design. If the cluster level panel cost is lower and the second interview cost is higher, the cluster-panel design is the most efficient. For inexpensive second interviews, the most efficient design is the observation-panel design. The latter domain includes our baseline case with $c_{12}^{\mathrm{I}} = 18$ and $c^{\mathrm{II}} = 3$.

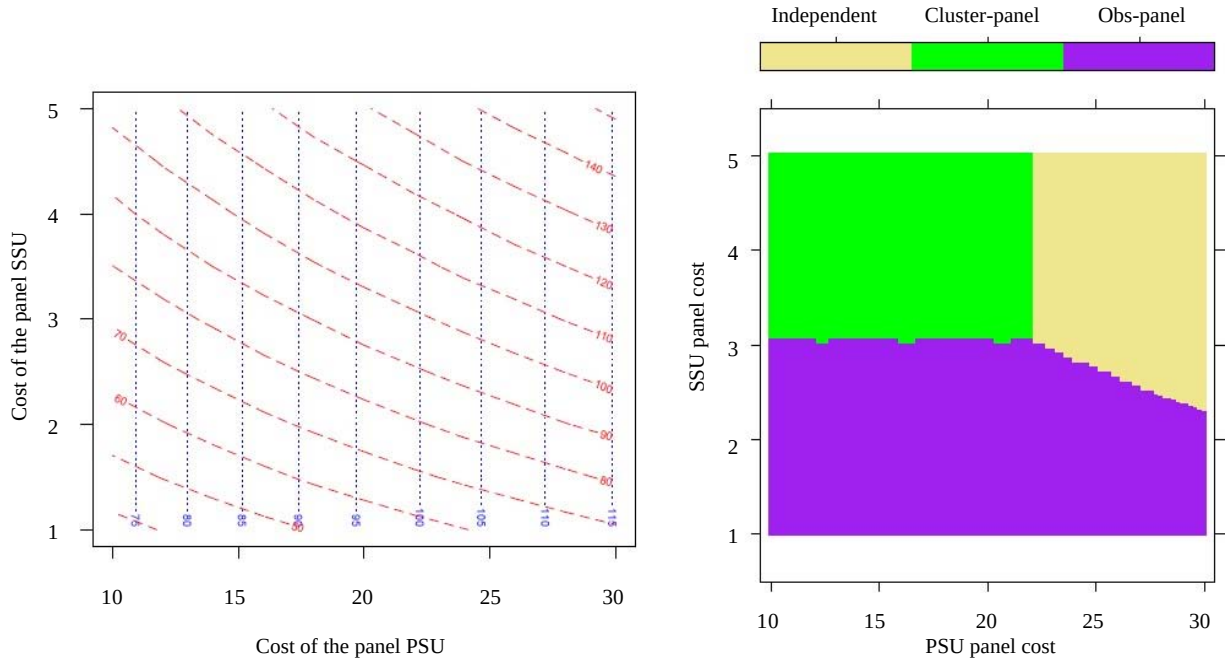


**Figure 1 Design variances as functions of the data collection costs $c_{12}^{\mathrm{I}}, c_{12}^{\mathrm{II}}$. Left: contour lines of $\mathrm{V}_{e,c}[d]$ (dotted) and $\mathrm{V}_{e,o}[d]$ (long dashed); $\mathrm{V}_{e,\iota} = 99.86$; right: domains of optimality of the three designs**

Figure 2 shows the changes in design variances associated with the changes in the autocorrelations $\rho^{I}$, $\rho^{II}$. The independent design variance is unaffected by these variations, and the cluster-panel design is unaffected by variations in $\rho^{II}$. The observation-panel design is more efficient for higher SSU autocorrelation, $\rho^{II} > 0.34$. Otherwise, the cluster-panel design provides lower variance.

Figure 3 investigates the impact of the cluster-level cost and autocorrelation on the choice of the design. The combinations of expensive second wave of data collection and low PSU autocorrelation in the upper left corner of the plot makes the independent design the most appealing one. Otherwise, the observation-panel design is the best one to use. Note that the contour lines for the cluster-panel and observation-panel designs are very close to one another, and differences in variances between the two designs are less than 2% in the whole parameter space of this plot.

Figure 4 investigates the impact of the observation-level cost and autocorrelation on the choice of the design. Neither the independent design nor the cluster-panel design variances are affected by variation of the parameters shown on this plot. The independent design variance is 99.86, while the cluster-panel design variance is 91.37, so the observation-panel design is compared to the latter only. High autocorrelations ($\rho^{II} \geq 0.6$) can justify very high cost of the second interview (up to fourfold compared to the first interview), but in the upper left corner of the plot corresponding to the low autocorrelations and high panel cost, the cluster-panel design performs better.

Figure 5 relates the design variances to the cluster-level costs of the survey. The horizontal axis is the cost in the first period, $c_1^I$, and the vertical axis is the additional cost of in the second period when the data are collected in a panel mode, $c_{12}^I - c_1^I$. The vertical axis is ignored for the independent design, as this parameter does not appear in the independent design. Also, by virtue of (4.1), $c_1^I = c_2^I$. The observation-panel design is uniformly better than the cluster-panel design for all parameter combinations on this graph, although the difference in variances does not exceed 2%. In the upper left corner, the additional cost of the panel mode of data collection is prohibitively high, and the independent design offers better performance.

Figure 6 shows the dependence of the most efficient design on the total budget of the survey and the cost of panel mode of data collection at the cluster level. For $C_0 > 10{,}000$, the observation-panel design performs better if $c_{12}^I < 22.7$, *i.e.*, if the additional cost of the panel mode of data collection at the cluster level does not exceed 127% of the initial cluster-level cost in the first wave. Interestingly, for some isolated parameter configurations in small surveys, the cluster-panel design can perform better than the observation-panel design that dominates the rest of the plot. The difference in design variances between the cluster-panel and observation-panel designs is less than 4% across all parameter combinations on this graph.

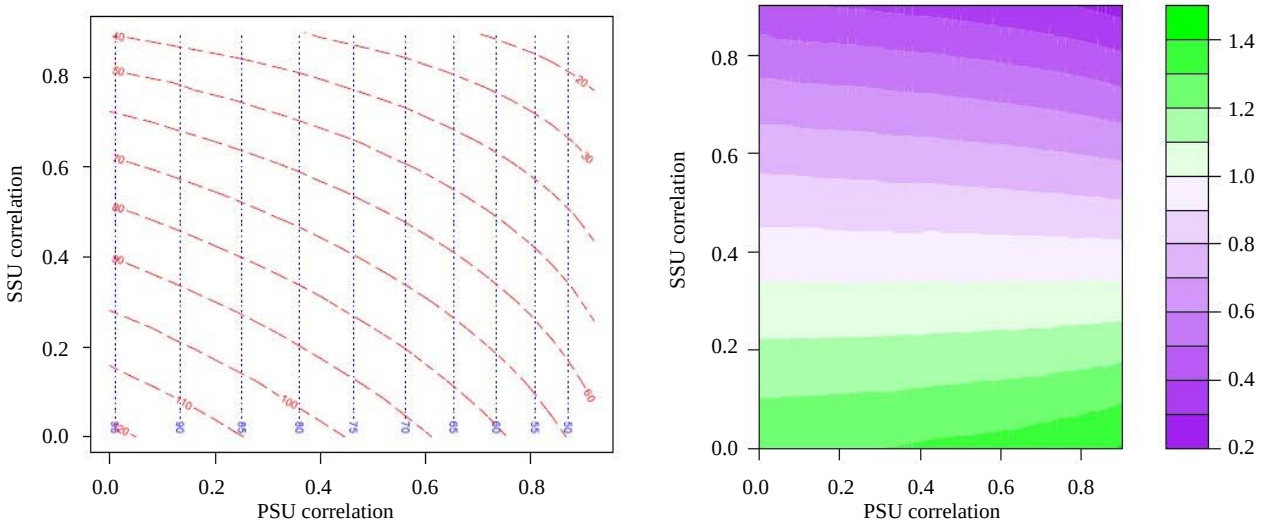


**Figure 2 Design variances as functions of the population correlations $\rho^{I}$, $\rho^{II}$. Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,\iota} = 99.86$; right: ratio $V_{e,o}[d] / V_{e,c}[d]$**

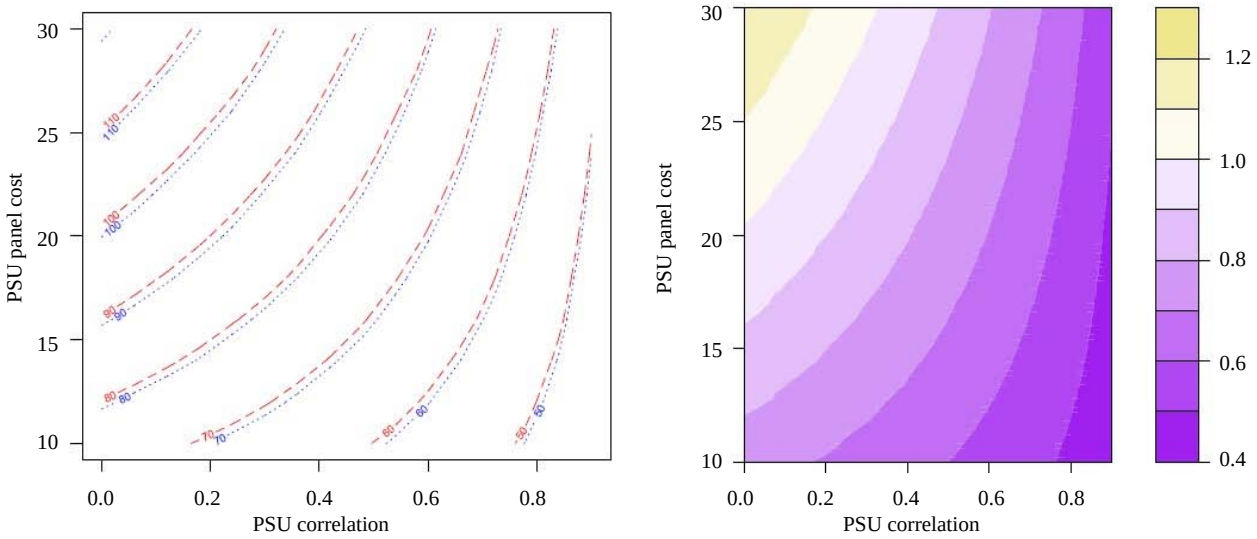


**Figure 3** Design variances as functions of the cluster-level autocorrelation $\rho^{I}$ and cost $c_{12}^{I}$. Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,\iota} = 99.86$; right: ratio $V_{e,o}[d] / V_{e,\iota}[d]$

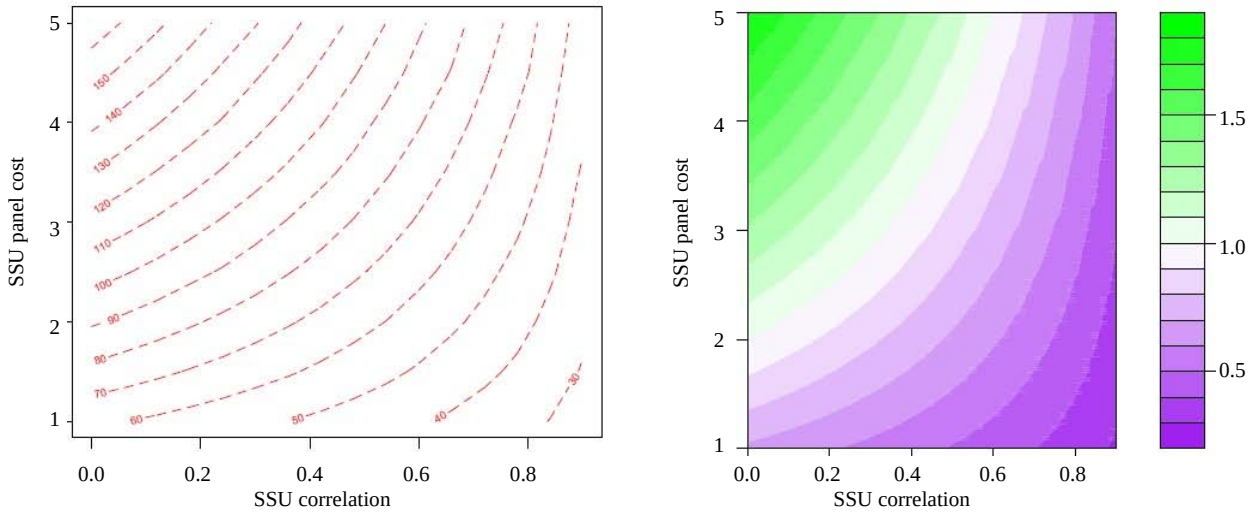


**Figure 4** Design variances as functions of the observation-level autocorrelation $\rho^{II}$ and cost $c_{12}^{II}$. Left: contour lines of $V_{e,o}[d]$ (long dashed); $V_{e,\iota} = 99.86$; $V_{e,c}[d] = 91.37$; right: ratio $V_{e,o}[d] / V_{e,c}[d]$

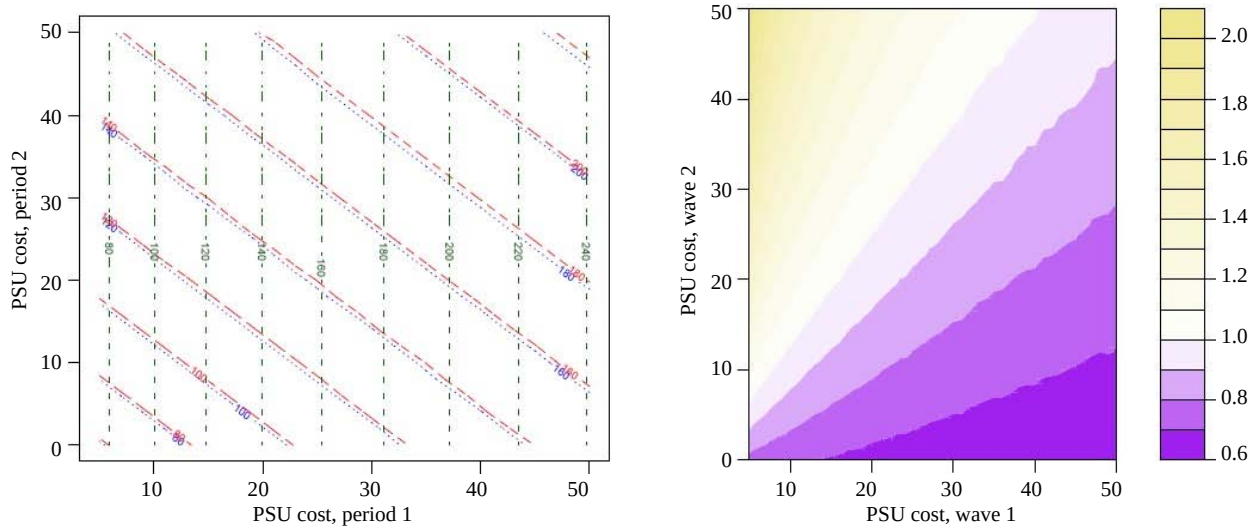


**Figure 5** Design variances as functions of the cluster level costs in the first wave, $c_1^{I}$, and in the second wave, $c_{12}^{I} - c_1^{I}$. Left: contour lines of $V_{e,c}[d]$ (dotted), $V_{e,o}[d]$ (long dashed) and $V_{e,\iota}[d]$ (dash-dotted); right: ratio $V_{e,o}[d] / V_{e,\iota}[d]$

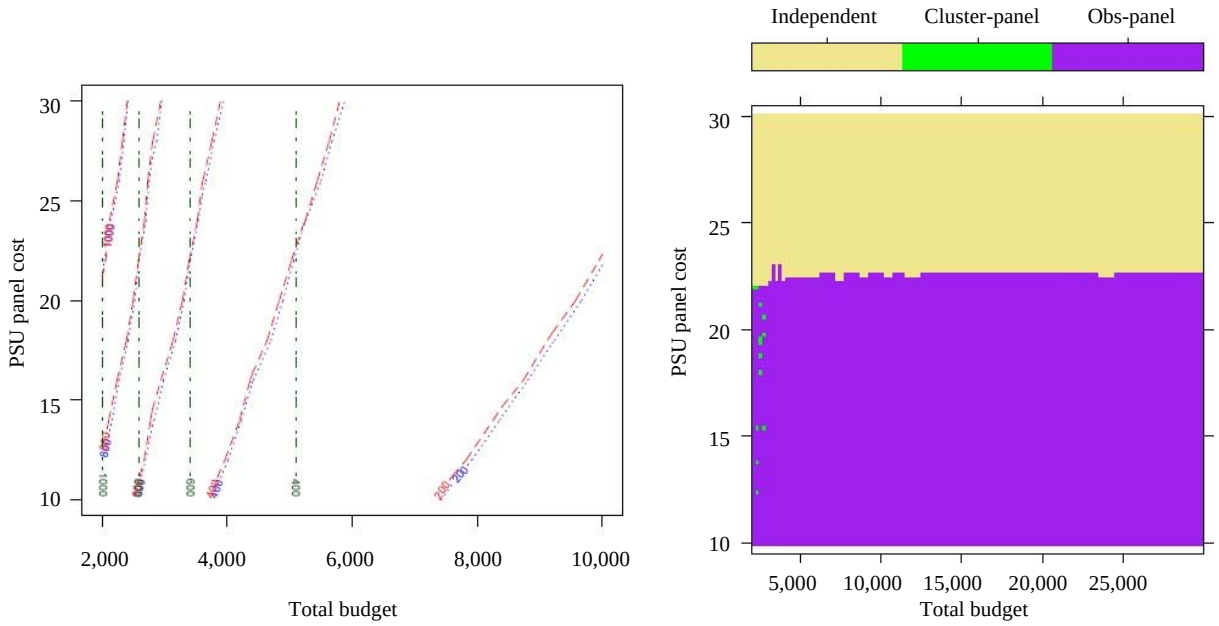


**Figure 6 Design variances as functions of the total budget $C_0$ and the PSU panel cost $c_{12}^{II}$. Left: contour lines of $V_{e,c}[d]$ (dotted), $V_{e,o}[d]$ (long dashed) and $V_{e,\iota}[d]$ (dash-dotted); right: domains of optimality of the three designs**

Overall, this numeric illustration shows that depending on the parameters of the population and costs of data collection, each of the three designs can be the most efficient one. Low correlations and high costs in the second wave tend to favor the independent design. Given that the initial six population parameters and five cost parameters may not be representative of many repeated surveys, a sensitivity analysis like the one performed here may be needed for any particular survey a statistician needs to design.

## 5. Survey design with multiple criteria

So far, our analysis was confined to estimation of the difference between the means in two waves of data collection of a single variable. Most large scale surveys are collected to study several characteristics, and to many users, the contemporaneous estimates are also of interest. To accommodate accuracy requirements associated with these different variables and different estimates, the survey designer must have several variances in mind when choosing the design to be implemented. This is a multicriterial optimization problem, and no single design will work best for all possible estimation problems. In the current context, the observation-panel design may give good estimates of the change when both PSU and SSU autocorrelations are high, but it may result in a small sample size if both PSUs and SSUs are expensive to follow up. Greater precision of the estimates for any single period could be obtained by switching to the cluster-panel or even independent designs.

Comparing different designs in this situation is possible with the standard microeconomic argument of utility maximization under budget constraints (Mas-Colell, Whinston and Green 1995). In the survey design context, the utility of the survey designer increases with the precision of the survey estimates, or equivalently decreases with survey variances. A simple functional form is given by Cobb-Douglas utility function:

$$U(\text{design}) = \mathrm{V}_{\text{design}}^{-\alpha_1}[\bar{y}_{1\cdot}]\,\mathrm{V}_{\text{design}}^{-\alpha_2}[\bar{y}_{2\cdot}]\,\mathrm{V}_{\text{design}}^{-\alpha_3}[d]. \quad (5.1)$$

Here, $\alpha_1$, $\alpha_2$ and $\alpha_3$ are positive constants describing the relative weights of the three design variances in decision-making process. Variances $\mathrm{V}[\bar{y}_1]$ and $\mathrm{V}[\bar{y}_2]$ in (5.2) are the variances of the means in cluster surveys given by (2.8). The variance of the difference estimator is (2.10), (2.12) or (2.14), depending on the design. The survey designer problem is then to maximize (5.1) subject to design-specific budget constraints (3.1), (3.4) or (3.6). Maximization is performed over the design parameters (mode of data collection, number of clusters in each time period, number of observations in each time period), given the characteristics of population (variances and autocorrelations) and the data collection process (costs).

Let us assume that the precision of each of the three estimates $\bar{y}_1$, $\bar{y}_2$ and $d$ is equally important to the decision maker, so $\alpha_1 = \alpha_2 = \alpha_3$. To have an objective function that is measured in the variance units and is on the same scale as variances, it will be convenient to define a multicriterial variance

$$V_{design} = (V_{design}[\bar{y}_{1\cdot\cdot}]\, V_{design}[\bar{y}_{2\cdot\cdot}]\, V_{design}[d])^{1/3}, \quad (5.2)$$

and express the optimization problem as minimization of this expression.

Analytic characterization of the design that optimizes (5.2) becomes quite cumbersome. Instead, we utilize a numeric illustration of the previous section to demonstrate how accounting for other design objectives affects the choice of the design. We should expect that for the designs with more expensive follow-ups ($c_{12}^{I} \geq c_1^{I} + c_2^{I}$, $c_{12}^{II} \geq c_1^{II} + c_2^{II}$), the simpler designs would be selected more often: the cluster-panel design may be preferred to the observation-panel design, and the independent design may be preferred to the cluster-panel design. For the baseline settings (4.1), we have

$$V_{e,\iota}[\bar{y}] = 49.93,\ V_{e,c}[\bar{y}] = 47.68,\ V_{e,o}[\bar{y}] = 61.69,$$

$$V_{e,\iota} = 62.91, \quad V_{e,c} = 59.23, \quad V_{e,o} = 70.02,$$

where the time indices of $y_{t\cdot\cdot}$ are omitted. The observation-panel design is rather inefficient in estimating the period-specific means as this design samples fewer units. Instead, the cluster-panel design is the most efficient one, closely followed by the independent design.

Figures 7-12 parallel Figure 1-6, respectively. Since the best design in terms of V is now the cluster-panel design, most of these plots show the preference toward this design. Figure 7 shows that when the variances of the contemporaneous means are taken into account, the simpler independent and cluster-panel designs are preferred for a greater fraction of parameter settings, and occupy a larger portion of the plot than in Figure 1. The point where the three designs are equivalent is $c_{12}^{I} = 20.6$, $c_{12}^{II} = 2.27$, closer to the origin than in Figure 1, in which only the variance of the difference was taken into account.

Figure 8 shows that the observation-panel design is only justified when both autocorrelations are higher than 0.6 (for the given values of population variances and costs). Recall that in Figure 2, the observation-panel design was preferred whenever $\rho^{II} > 0.34$, with little dependence on $\rho^{I}$.

Figure 9 shows how the PSU-level correlations and costs affect the choice of the design. The observation-panel design is less efficient than the cluster-panel design for all combinations of parameters in this plot. Hence, the choice of the design is between the independent and the cluster-panel designs. Naturally, if the data collection in the panel mode is expensive, the independent design is preferred to the cluster-panel design. Interestingly, the preference towards a particular design is not monotone in $\rho_{12}^{I}$. With values $\rho_{12}^{I} > 0.7$, the $V[d]$ component in (5.2) produces designs with so few clusters that $V[\bar{y}]$ suffers notably enough to hurt the whole objective function. At that value of panel autocorrelation, the maximum panel cost at which the cluster-panel design is still the most efficient one is $c_{12}^{I} = 24.4$, *i.e.*, the cluster-level cost in the second wave is 44% higher than in the first wave.

Figure 10 shows that the higher autocorrelation of the SSU measurements may justify modest extra cost associated with data collection. The highest cost for which the observation-panel design is still the most efficient one is $c_{12}^{II} = 2.75$ with $\rho^{II} = 0.78$; *i.e.*, the cost of the second interview can be 75% more than the cost of the first interview.

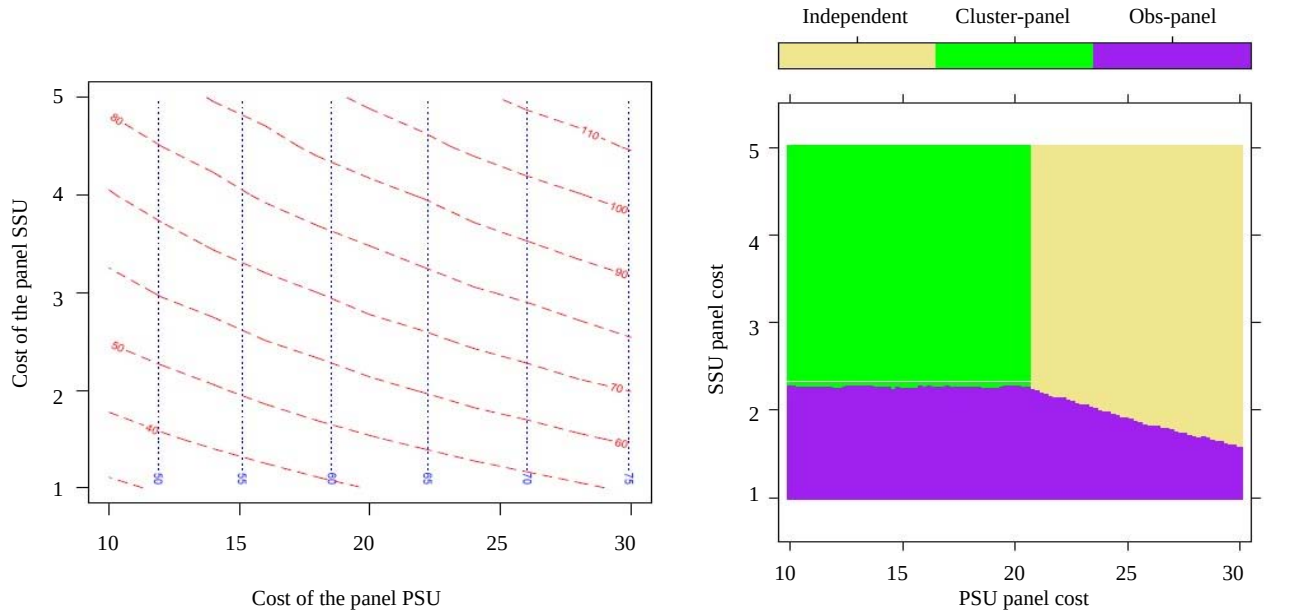


**Figure 7 Design variances as functions of the data collection costs $c_{12}^{I}, c_{12}^{II}$. Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; right: domains of optimality of the three designs**

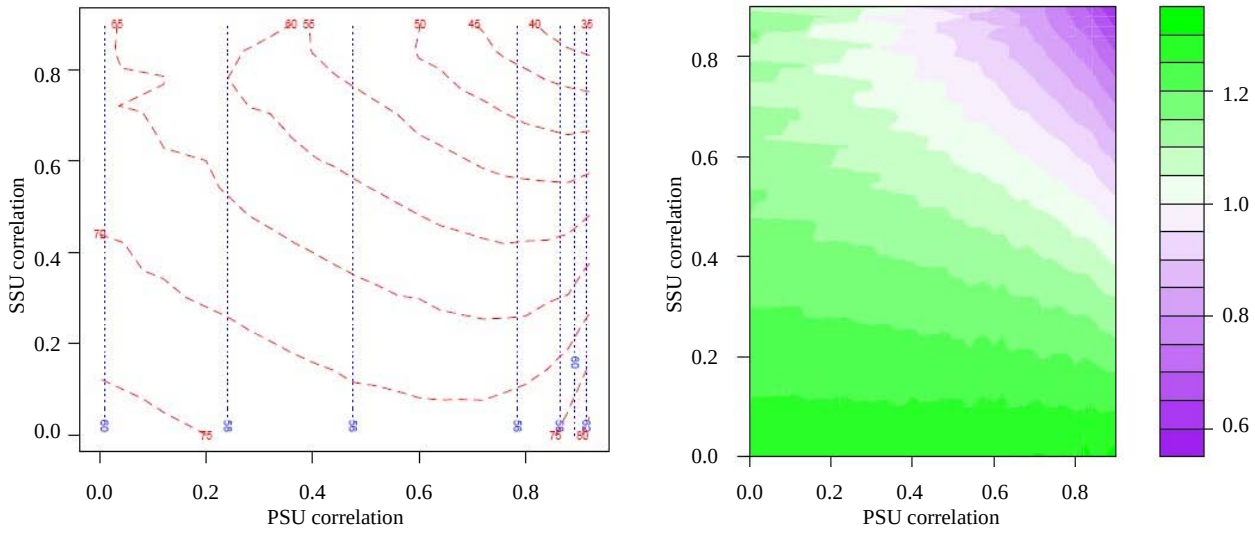


**Figure 8** Design variances as functions of the autocorrelations $\rho^{I}, \rho^{II}$. Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; right: ratio $V_{e,o}/V_{e,c}$

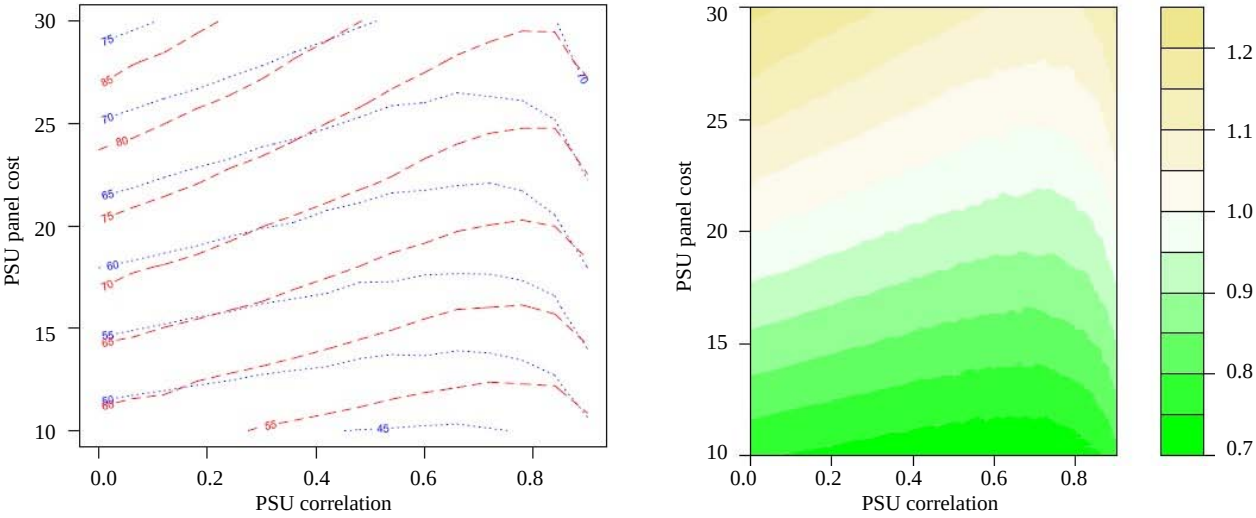


**Figure 9** Design variances as functions of the cluster-level autocorrelation $\rho^{I}$ and cost $c_{12}^{I}$. Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; right: ratio $V_{e,c}/V_{e,\iota}$

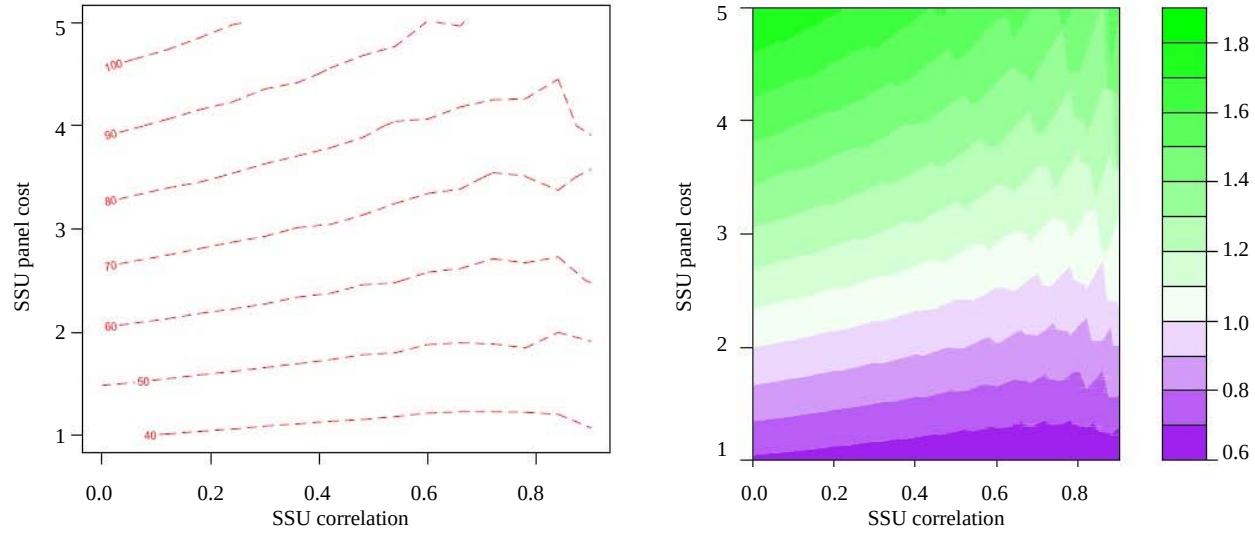


**Figure 10** Design variances as functions of the observation-level autocorrelation $\rho^{II}$ and cost $c_{12}^{II}$. Left: contour lines of $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; $V_{e,c} = 59.23$; right: ratio $V_{e,o}/V_{e,c}$

Figure 11 parallels Figure 5. The left panel shows that the observation-panel design is less efficient than the cluster-panel design. The right panel shows that if the cluster-level cost of the second wave exceeds the cluster-level cost of the first wave by more than 15 units, the independent design delivers better efficiency than the cluster-panel design.

Finally, Figure 12 shows the variances as functions of the total survey budget and the cost of the panel mode of data collection. There is very little dependence on $C_0$ in the plot, and the independent design is preferred if the panel mode is too expensive, namely, when the cluster-level cost in the second cost exceeds 107% of that in the first wave.

As it was conjectured in the beginning of this section, incorporation of the variances of the contemporaneous means into the design optimization objective function shifted the preferences of the survey designer towards simpler designs that can sample a greater number of the ultimate observation units. The observation-panel design now only makes sense when both the PSU and SSU autocorrelations are high, and the panel costs are reasonably low. Moreover, the cluster-panel design is generally justified only if there is an economy in cluster-level cost in the second wave of the survey.

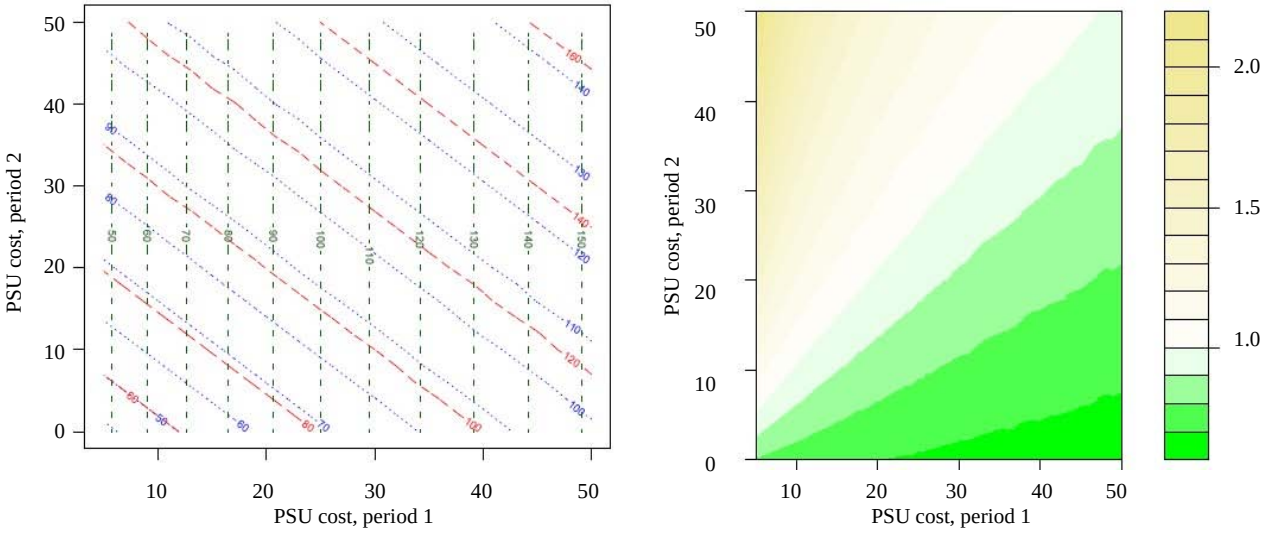


**Figure 11 Design variances as functions of the data collection costs $c_1^{\mathrm{I}}, c_{12}^{\mathrm{I}}$. Left: contour lines of $\mathrm{V}_{e,c}$ (dotted), $\mathrm{V}_{e,o}$ (long dashed) and $\mathrm{V}_{e,\iota}$ (dash-dotted); right: ratio $\mathrm{V}_{e,c}/\mathrm{V}_{e,\iota}$**

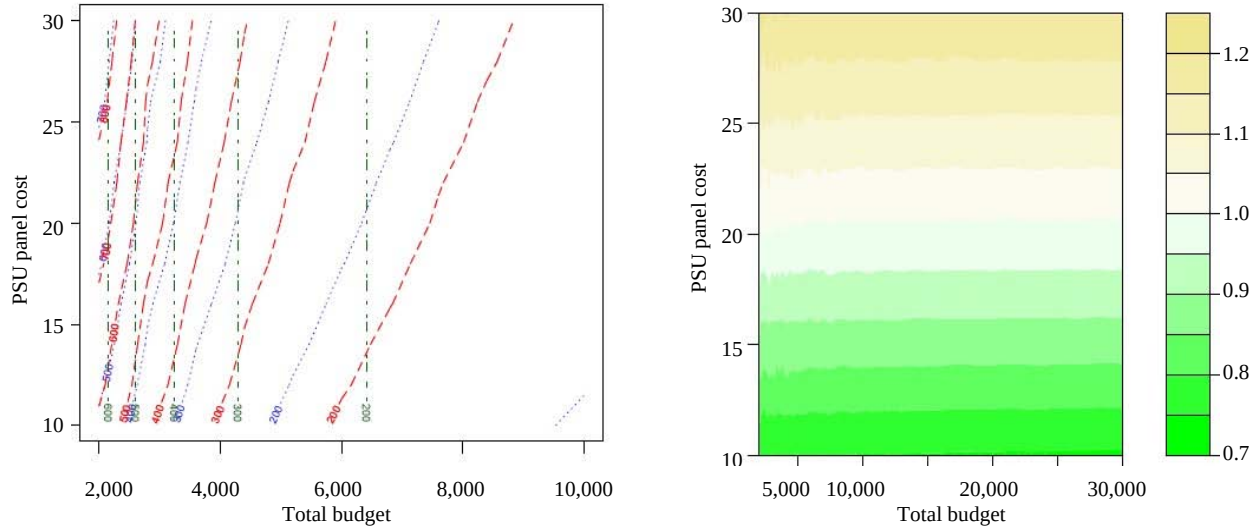


**Figure 12 Design variances as functions of the total budget $C_0$ and the PSU panel cost $c_{12}^{\mathrm{II}}$. Left: contour lines of $\mathrm{V}_{e,c}$ (dotted), $\mathrm{V}_{e,o}$ (long dashed) and $\mathrm{V}_{e,\iota}$ (dash-dotted); right: domains of optimality of the three designs**

## 6. Extensions to multiple waves

If the survey to be designed will have more than two waves of data collection, the survey designer may be able to extend the framework of the utility maximization problem (5.1), with the following considerations in mind.

1. A greater number of targets of inference. Possible variances that the survey designer may need to take into account can now include: contemporaneous variances $V[\bar{y}_1]$, $V[\bar{y}_2]$, …, $V[\bar{y}_T]$; consecutive differences $V[\bar{y}_2 - \bar{y}_1]$, …, $V[\bar{y}_T - \bar{y}_{T-1}]$ or composite/GLS estimators of the change between two adjacent periods of time; other contrasts $V[\sum_t c_t \bar{y}_t], \sum c_t = 0$; variance of the linear growth rates from regression of $\bar{y}_t$ on $t$, estimated by OLS or GLS; *etc.*
2. A possibility of discounting. In economics, it is customary to specify the budget constraints that look into the future in the form of $\sum_t x_t \delta^t$ where $x_t$ is the amount spent in time $t$, and $\delta < 1$ is the discount factor associated with interest rates. Discounting may also be relevant for the utility function, and design variances farther in the future may have lower weights in the optimization problem.
3. Unknown functional forms of the time-series processes associated with the variable of interest. The survey designer needs to have a good idea about the covariance structure of the time series of both individual observations and cluster means. It is likely that the results will be sensitive to the choice of the particular model. In the current analysis, the issue is ameliorated, as it suffices to have a single correlation parameter for each level. The survey designer may have to introduce more parameters into the model, and correspondingly study sensitivity of the design choice with respect to these parameters.

The complexity of the problem, as outlined above, can grow out of control very quickly. We thus abstain from a more detailed treatment of it in this paper.

## 7. Discussion

This paper has analyzed different options for implementation of repeated cluster surveys. We have provided analytical expression for design variances of the simple difference estimator for three popular designs (the independent, the cluster-panel and the observation-panel designs). We have also derived the optimal sample sizes for estimation of the difference between two waves of data collection.

The sample designer who knows that the characteristic of interest is going to have some degree of persistence over time will likely choose one of the panel designs, provided that the costs of re-visiting the clusters and/or observation units are not prohibitively high. Analytical comparison is possible between the independent and the observation-panel designs, and is given by Proposition 7. It is worth noting that the design variance of the difference is $O(C_0^{-1})$ for both the independent design and the observation-panel design, and is $O(C_0^{-1/2})$ for the cluster-panel design, where $C_0$ is the total budget of the survey. Hence the cluster-panel design is only viable for smaller surveys, while the large scale surveys will likely have either the independent or the observation-panel format.

The cost structure considered in Section 3 is rather simplistic. For instance, the second stage costs in the second time period may differ across individuals sampled from the new or from the reused clusters. 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. Our original motivation was to consider situations in which the SSU panel cost is higher than twice the cost of individual interviews. However, as suggested by one of the referees, this cost may be lower if the follow-up interviews are performed in cheaper mode, such as a phone interview or a self-administered mail survey instead of a personal interview. If this is the case, the observation-panel design is apparently the most cost-efficient of the three designs.

The population structure is also an oversimplification. The clusters are assumed to be of balanced unchanging sizes. No units leave the population, and no new units appear. These assumptions are quite restrictive for many practical situations. If the population changes between two waves of data collection, the sample designer would want to include new clusters at the second wave, using the algorithms of Ernst (1999). The new clusters are placed into a separate stratum, and a clustered sample is taken from that stratum. In NHIS, this is implemented by "permit" frame. 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 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). Arguably, both non-response and loss of coverage can be added to the current framework as sources of bias, leading to optimization of the mean squared total survey error rather than the design variance. Convincing models of such biases may be difficult to formulate, however.

Another issue that would arise with clusters of different sizes is that of the greater range of applicable designs. In this paper, we assumed SRSWOR at both stages. Other designs, such as sampling with probability proportional to size (PPS), can be used instead. For designs other than SRS, the Horvitz-Thompson estimator and its variance (Särndal, Swensson and Wretman 1992, Thompson 1997) would need to be used. The analytical derivations become unwieldy, although practical numerical demonstrations similar to our Sections 4 and 5 can still be implemented. If cluster sizes change over time, obtaining the optimal design becomes a moving target, and designs optimal for the "old" measures of size will lose their efficiency with the "new" measures of size.

In earlier drafts of this paper, we analyzed intermediate designs where a non-trivial fraction of the units are retained, and other units are sampled independently. The problem can then be viewed as variance minimization subject to inequality constraints on the degree of the overlap $0 \le \pi^{\mathrm{I}} \le 1$, $0 \le \pi^{\mathrm{II}} \le 1$. The general theory of non-linear constrained optimization ensures that as long as the variance of the population mean change $D$ is monotone in $\pi^{\mathrm{I}}$ and $\pi^{\mathrm{II}}$, the optimum will be achieved in one of the vertices of the parameter space. This justifies our interest in the three designs considered in the paper. They correspond to the vertices of the parameter space: $(0, 0)$, $(1, 0)$ and $(1, 1)$ for the independent, cluster-panel and observation-panel designs, respectively. The point $(0, 1)$ corresponds to an impossible design with complete overlap of the individual units with no overlap of the clusters. Cumbersome derivations show that it is possible to satisfy the first order conditions in some intermediate cases, too, but they correspond to local maxima of the variance. While these results may also be of interest (in the sense of providing an upper bound on the design variances), we did not consider them in the paper. In the more complicated cases of the multicriterial optimization of Section 5, monotonicity does not necessarily hold, and other designs beside the three extreme cases considered in the paper may lead to the optimal values of the objective function (5.2).

Conditions of equal variances (2.9) can be relaxed at the price of producing substantially more complicated expressions. If the sample sizes are fixed between the two occasions, then the following changes will be necessary in all relevant formulas. In the expressions that do not involve autocorrelations,

$$2S_b^2 \mapsto S_{1b}^2 + S_{2b}^2, \quad 2S_w^2 \mapsto \overline{S}_{1w}^2 + \overline{S}_{2w}^2, \qquad (7.1)$$

while in the expressions that do involve autocorrelations,

$$2(1-\rho^{\mathrm{I}})S_b^2 \mapsto S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b},$$
$$2S_w^2(1-\rho^{\mathrm{II}}) \mapsto \overline{S}_{1w}^2 + \overline{S}_{2w}^2 - 2\rho^{\mathrm{I}} \overline{S}_{1w} \overline{S}_{2w}. \qquad (7.2)$$

Qualitatively, the results will be the same.

The multicriterial framework of Section 5 allows for different importance weights to be given to different variances of interest. Relatively larger values of $\alpha_1, \alpha_2$ correspond to the greater importance of the contemporaneous means, while larger values of $\alpha_3$ correspond to the greater importance of the change estimate. The original problem of optimizing the design for $\mathrm{V}[d]$ can be considered within the context of (5.1) by setting $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 1$. This framework can also be expanded to include designs aimed at measuring several variables. An additional challenge of such a setup is that the autocorrelations may differ across different variables. Some individual characteristics are constant over time (race, gender); others change slowly (housing, expenditure, political preferences), yet others may change faster (income or behavior).

This paper dealt with three designs and a specific estimator of change: the difference in the two estimates of the mean in two periods of time. Other options for either designs or estimators are also available. For instance, in rotation designs, a fraction of the first wave units is retained, and some new units are recruited. For such designs, composite estimation (Hansen *et al.* 1953, Patterson 1950, Rao and Graham 1964, Wolter 2007) that weighs differently the contributions of the independent units (those retired from the sample after the first wave, and those newly recruited for the second wave) and the contributions of the panel units (used in both waves) would result in more efficient estimates. Generally, motivation for such designs comes from non-sampling considerations, such as decrease of the response burden and deterioration of the sample representativeness of population due to the population change. These considerations can be accounted for in either the cost model (*e.g.*, a greater number of callbacks required to convince a unit to respond), or the total survey error model (by introducing the non-response or undercoverage bias, and considering mean squared error rather than the design variance of an estimate).

## Acknowledgements

The authors are grateful to Chris Skinner and John Eltinge for helpful discussions, to William Kalsbeek for suggestions at the early stages of the paper, and to the associate editor and two referees for their comments. Nash Herndon and Oksana Loginova provided editorial improvements. Partial 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. The authors are also grateful to the participants of the Joint Statistical Meetings

(2005) and the XXIII International Methodology Symposium of Statistics Canada (2007) for helpful comments.

## Appendix

Expectations, variances and covariances in the proofs below are with respect to the corresponding designs. The first stage of selection will be denoted with a superscript I. The second stage of selection will be denoted with a superscript II.

*Proof of Proposition* 2. Let us denote the sample of the PSUs by $\mathcal{S}^{\mathrm{I}}$, the sample of SSUs in the first period by $\mathcal{S}^{\mathrm{II}}_{i1}$, and the sample of SSUs in the second period by $\mathcal{S}^{\mathrm{II}}_{i2}$. Then

$$d = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} = \frac{1}{mn}\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij} - \sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right).$$

Denoting the expectations with respect to the first stage as $\mathrm{E}_{\mathrm{I}}$, and those with respect to the second stage as $\mathrm{E}_{\mathrm{II}}$, we have the design variance of $d$ equal to

$$\begin{aligned}
\mathrm{V}[d] &= \mathrm{E}_{\mathrm{I}}\mathrm{V}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] + \mathrm{V}_{\mathrm{I}}\mathrm{E}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij} - \sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right]\right\} \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{E}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij} - \sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right]\right\} \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij}\right] + \mathrm{V}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right]\right\} \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}} m\bar{Y}_{2i\cdot} - m\bar{Y}_{1i\cdot}\right] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(1-\frac{m}{M}\right)mS^2_{2wi} + \left(1-\frac{m}{M}\right)mS^2_{1wi}\right] \\
&\quad + \frac{1}{n^2}\left(1-\frac{n}{N}\right)n(S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \frac{1}{m^2n^2}nm\left(1-\frac{m}{M}\right)m(S^2_{2w} + S^2_{1w}) \\
&\quad + \frac{1}{n}\left(1-\frac{n}{N}\right)(S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \left(1-\frac{n}{N}\right)\frac{2S^2_b(1-\rho^{\mathrm{I}})}{n} + \left(1-\frac{m}{M}\right)\frac{2S^2_w}{mn},
\end{aligned}$$

where the last equality assumes symmetric conditions (2.9).

*Proof of Proposition* 3. Let us denote the sample of the PSUs by $\mathcal{S}^{\mathrm{I}}$, and the sample of SSUs, by $\mathcal{S}^{\mathrm{II}}_i$. Then

$$d = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} = \frac{1}{mn}\sum_{i\in\mathcal{S}^{\mathrm{I}}}\sum_{j\in\mathcal{S}^{\mathrm{II}}_i}(y_{2ij} - y_{1ij}).$$

Denoting the expectations with respect to the first stage as $\mathrm{E}_{\mathrm{I}}$, and those with respect to the second stage as $\mathrm{E}_{\mathrm{II}}$, we have the design variance of $d$ equal to

$$\begin{aligned}
\mathrm{V}[d] &= \mathrm{E}_{\mathrm{I}}\mathrm{V}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] + \mathrm{V}_{\mathrm{I}}\mathrm{E}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V}_{\mathrm{II}}\sum_{j\in\mathcal{S}^{\mathrm{II}}_i}(y_{2ij} - y_{1ij})\right] \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{E}_{\mathrm{II}}\sum_{j\in\mathcal{S}^{\mathrm{II}}_i}(y_{2ij} - y_{1ij})\right] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}m\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(1-\frac{m}{M}\right)(S^2_{2wi}+S^2_{1wi}-2S_{2wi}S_{1wi}\rho^{\mathrm{II}})\right] \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}} m(\bar{Y}_{2ij} - \bar{Y}_{1ij})\right] \\
&= \frac{1}{mn^2}n\left(1-\frac{m}{M}\right)(\bar{S}^2_{1w} + \bar{S}^2_{2w} - 2\rho^{\mathrm{II}}\bar{S}_{1w}\bar{S}_{2w}) \\
&\quad + \frac{1}{n^2}n\left(1-\frac{n}{N}\right)(S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \left(1-\frac{n}{N}\right)\frac{S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}}{n} \\
&\quad + \left(1-\frac{m}{M}\right)\frac{\bar{S}^2_{1w} + \bar{S}^2_{2w} - 2\rho^{\mathrm{II}}\bar{S}_{1w}\bar{S}_{2w}}{mn} \\
&= 2\left(1-\frac{n}{N}\right)\frac{S^2_b(1-\rho^{\mathrm{I}})}{n} \\
&\quad + 2\left(1-\frac{m}{M}\right)\frac{S^2_w(1-\rho^{\mathrm{II}})}{mn},
\end{aligned}$$

with the last equality holding under the symmetry conditions.

*Proof of Proposition* 4. The Lagrangian function of minimizing (2.11) subject to constraint (3.1) is

$$\begin{aligned}
L(n_1, m_1, n_2, m_2, \lambda) = &\left(1-\frac{n_1}{N}\right)\frac{S^2_b}{n_1} + \left(1-\frac{n_2}{N}\right)\frac{S^2_b}{n_2} \\
&+ \left(1-\frac{m_1}{M}\right)\frac{\bar{S}^2_w}{n_1m_1} + \left(1-\frac{m_2}{M}\right)\frac{\bar{S}^2_w}{n_2m_2} \\
&- \lambda(c^{\mathrm{I}}_1n_1 + c^{\mathrm{II}}_1n_1m_1 + c^{\mathrm{I}}_2n_2 + c^{\mathrm{II}}_2n_2m_2 - C_0).
\end{aligned}$$

Working through the first order conditions of this Lagrangian function leads to

$$-\lambda = \frac{m_1 S_b^2 + \left(1 - \frac{m_1}{M}\right)\bar{S}_w^2}{n_1^2 m_1 (c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1)} = \frac{m_2 S_b^2 + \left(1 - \frac{m_2}{M}\right)\bar{S}_w^2}{n_2^2 m_2 (c_2^{\mathrm{I}} + c_1^{\mathrm{II}} m_2)}$$
$$= \frac{\bar{S}_w^2}{m_1^2 n_1^2 c_1^{\mathrm{II}}} = \frac{\bar{S}_w^2}{m_2^2 n_2^2 c_2^{\mathrm{II}}}.$$

Utilizing these conditions, we have

$$m^2 n^2 c^{\mathrm{II}} \left[ m S_b^2 + \left(1 - \frac{m}{M}\right)\bar{S}_w^2 \right] = n^2 m (c^{\mathrm{I}} + c^{\mathrm{II}} m)\bar{S}_w^2,$$

which can be written as

$$0 = (c^{\mathrm{I}} + c^{\mathrm{II}} m)\bar{S}_w^2 - m c^{\mathrm{II}} \left[ m S_b^2 + \left(1 - \frac{m}{M}\right)\bar{S}_w^2 \right]$$
$$= (c^{\mathrm{I}} + c^{\mathrm{II}} m) M \bar{S}_w^2 - m c^{\mathrm{II}} [M m S_b^2 + (M - m)\bar{S}_w^2]$$
$$= c^{\mathrm{I}} M \bar{S}_w^2 + m c^{\mathrm{II}} M \bar{S}_w^2 - m^2 c^{\mathrm{II}} M S_b^2 - m c^{\mathrm{II}} M \bar{S}_w^2 + m^2 c^{\mathrm{II}} \bar{S}_w^2$$
$$= c^{\mathrm{I}} M \bar{S}_w^2 + m^2 c^{\mathrm{II}} (\bar{S}_w^2 - M S_b^2).$$

Hence,

$$m = \sqrt{\frac{c^{\mathrm{I}}}{c^{\mathrm{II}}} \frac{\bar{S}_w^2}{S_b^2 - \bar{S}_w^2 / M}}.$$

From the survey budget (3.1), the number of clusters is found to be

$$n = \frac{C_0}{2(c^{\mathrm{I}} + m c^{\mathrm{II}})} = \frac{C_0}{2\{c^{\mathrm{I}} + [c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}}.$$

Plugging these expressions into (2.11) and using the equality relations (2.9), we obtain the variance of the estimator as

$$\mathrm{V}_{e,\iota}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2}{mn}$$
$$= 2\left[1 - \frac{C_0}{2(c^{\mathrm{I}} + m c^{\mathrm{II}})N}\right]\frac{2(c^{\mathrm{I}} + m c^{\mathrm{II}})S_b^2}{C_0}$$
$$+ 4\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2 (c^{\mathrm{I}} + m c^{\mathrm{II}})}{m C_0}$$
$$= 2\left[\frac{2(c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0} - \frac{1}{N}\right]S_b^2$$
$$+ 4\left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{S}_w^2 (c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0}$$
$$= \frac{4(c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0}\left[S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2\right] - \frac{2}{N}S_b^2$$
$$= \frac{4\left[c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)}\right]}{C_0}$$
$$\times \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}\frac{S_b^2 - \bar{S}_w^2 / M}{\bar{S}_w^2}} - \frac{1}{M}\right)\bar{S}_w^2\right] - \frac{2}{N}S_b^2.$$

*Proof of Proposition* 5. The Lagrangian function of minimizing (2.13) subject to constraint (3.4) is

$$L(n, m_1, m_2, \lambda) = 2\left(1 - \frac{n}{N}\right)\frac{(1 - \rho^{\mathrm{I}})S_b^2}{n}$$
$$+ \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n m_1} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n m_2}$$
$$- \lambda(c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2 - C_0).$$

The first order conditions are:

$$\frac{\partial L}{\partial n} = -2\frac{(1 - \rho^{\mathrm{I}})S_b^2}{n^2} - \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n^2 m_1}$$
$$- \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n^2 m_2} - \lambda(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2),$$
$$\frac{\partial L}{\partial m_1} = -\frac{\bar{S}_{1w}^2}{n m_1^2} - \lambda c_1^{\mathrm{II}} n,$$
$$\frac{\partial L}{\partial m_2} = -\frac{\bar{S}_{2w}^2}{n m_2^2} - \lambda c_2^{\mathrm{II}} n,$$
$$\frac{\partial L}{\partial \lambda} = c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2 - C_0 = 0.$$

Expressing $-\lambda n$ from these conditions, one obtains:

$$-\lambda n = \frac{\bar{S}_{1w}^2}{m_1^2 n c_1^{\mathrm{II}}} = \frac{\bar{S}_{2w}^2}{m_2^2 n c_2^{\mathrm{II}}} = 2(1 - \rho^{\mathrm{I}})\frac{S_b^2}{C_0}$$
$$+ \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{m_1 C_0} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{m_2 C_0}.$$

Then

$$\frac{1}{m_2} = \frac{1}{m_1}\sqrt{\frac{c_2^{\mathrm{II}}}{c_1^{\mathrm{II}}}\frac{\bar{S}_{1w}^2}{\bar{S}_{2w}^2}} \equiv \frac{1}{\kappa m_1},$$
$$\frac{1}{m_1^2}\frac{(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + \kappa c_2^{\mathrm{II}} m_1)\bar{S}_{1w}^2}{c_1^{\mathrm{II}}} = 2(1 - \rho^{\mathrm{I}})S_b^2$$
$$+ \left(\frac{1}{m_1} - \frac{1}{M}\right)\bar{S}_{1w}^2 + \left(\frac{1}{\kappa m_1} - \frac{1}{\kappa M}\right)\bar{S}_{2w}^2,$$
$$0 = [2(1 - \rho^{\mathrm{I}})S_b^2 \kappa c_1^{\mathrm{II}} - \bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} / M - \bar{S}_{2w}^2 c_1^{\mathrm{II}} / M]m_1^2$$
$$+ [\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} + \bar{S}_{2w}^2 c_1^{\mathrm{II}} - c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2]m_1 - c_{12}^{\mathrm{I}} \bar{S}_{1w}^2 \kappa,$$
$$D = [\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} + \bar{S}_{2w}^2 c_1^{\mathrm{II}} - c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2]^2$$
$$+ 4[2(1 - \rho^{\mathrm{I}})S_b^2 \kappa - \bar{S}_{1w}^2 \kappa / M - \bar{S}_{2w}^2 / M]c_1^{\mathrm{II}} c_{12}^{\mathrm{I}} \bar{S}_{1w}^2 \geq 0,$$
$$m_1 = \frac{c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa + \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2 - \bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} - \bar{S}_{2w}^2 c_1^{\mathrm{II}} \pm \sqrt{D}}{4(1 - \rho^{\mathrm{I}})S_b^2 \kappa c_1^{\mathrm{II}} - 2\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} / M - 2\bar{S}_{2w}^2 c_1^{\mathrm{II}} / M}.$$

The solution with $-\sqrt{D}$ leads to a negative value of $m_1$, and must be discarded.

The remaining design characteristics are

$$m_2 = \kappa m_1, \quad n = \frac{C_0}{c_{12}^{\mathrm{I}} + m_1 c_1^{\mathrm{II}} + m_2 c_1^{\mathrm{II}}}, \quad \kappa = \sqrt{\frac{c_1^{\mathrm{II}}}{c_2^{\mathrm{II}}}\frac{\bar{S}_{2w}^2}{\bar{S}_{1w}^2}}.$$

The variance of the difference estimator can be found using (2.15).

Under symmetric conditions, $\kappa = 1$, and

$$D = 4[2(1-\rho^{I})S_b^2 - 2\bar{S}_w^2/M]\, c^{II}c_{12}^{I}\bar{S}_w^2$$

is non-negative unless the expression in the square brackets is negative (which can only happen when $\rho^{I}$ is large and $M$ is small. In that case, a corner solution $m = M$ is realized). Furthermore,

$$m = m_1 = m_2 = \sqrt{\frac{\bar{S}_w^2 c_{12}^{I}}{2[(1-\rho^{I})S_b^2 - \bar{S}_w^2/M]c^{II}}},$$

$$n = \frac{C_0}{c_{12}^{I} + 2mc^{II}} = \frac{C_0}{c_{12}^{I} + \sqrt{\dfrac{2\bar{S}_w^2 c_{12}^{I} c^{II}}{(1-\rho^{I})S_b^2 - \bar{S}_w^2/M}}},$$

$$\begin{aligned}
&\mathrm{V}_{e,o}[d] \\
&= 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{nm} \\
&= \frac{2}{n}\left[(1-\rho^{I})S_b^2 + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{m}\right] - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}(c_{12}^{I} + 2mc^{II}) \\
&\times\left[(1-\rho^{I})S_b^2 + 2\left(\frac{1}{m}-\frac{1}{M}\right)(1-\rho^{II})\bar{S}_w^2\right] - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}(c_{12}^{I} + 2mc^{II}) \\
&\times\left[(1-\rho^{I})S_b^2 + \frac{2}{m}(1-\rho^{II})\bar{S}_w^2 - \frac{2}{M}(1-\rho^{II})\bar{S}_w^2\right] - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{c_{12}^{I}(1-\rho^{I})S_b^2 + 2(1-\rho^{II})\bar{S}_w^2\left[2c^{II} - \frac{c_{12}^{I}}{M}\right]\right. \\
&\quad \left. + \frac{2}{m}c_{12}^{I}(1-\rho^{II})\bar{S}_w^2 + 2mc^{II}\left[(1-\rho^{I})S_b^2 - \frac{2}{M}(1-\rho^{II})\bar{S}_w^2\right]\right\} \\
&\quad - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{c_{12}^{I}(1-\rho^{I})S_b^2 + 2(1-\rho^{II})\bar{S}_w^2\left[2c^{II} - \frac{c_{12}^{I}}{M}\right]\right. \\
&\quad + 2(1-\rho^{II})\sqrt{2[(1-\rho^{I})S_b^2 - \bar{S}_w^2/M]\bar{S}_w^2 c^{II}c_{12}^{I}} \\
&\quad \left. + \sqrt{\frac{2\bar{S}_w^2 c_{12}^{I}c^{II}}{(1-\rho^{I})S_b^2 - \bar{S}_w^2/M}}\left[(1-\rho^{I})S_b^2 - \frac{2}{M}(1-\rho^{II})\bar{S}_w^2\right]\right\} \\
&\quad - \frac{2(1-\rho^{I})S_b^2}{N}.
\end{aligned}$$

*Proof of Proposition* 6. The Lagrangian function of minimizing (2.15) subject to constraint (3.6) is

$$\begin{aligned}
L(n, m, \lambda) &= 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{nm} \\
&\quad - \lambda(c_{12}^{I}n + c_{12}^{II}nm - C_0).
\end{aligned}$$

The first order conditions are:

$$\begin{aligned}
0 &= \frac{\partial L}{\partial n} = -2\frac{(1-\rho^{I})S_b^2}{n^2} \\
&\quad - 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{n^2 m} - \lambda(c_{12}^{I} + c_{12}^{II}m), \\
0 &= \frac{\partial L}{\partial m} = -2\frac{(1-\rho^{II})\bar{S}_w^2}{nm^2} - \lambda c_{12}^{II}n, \\
0 &= \frac{\partial L}{\partial \lambda} = c_{12}^{I}n + c_{12}^{II}nm - C_0.
\end{aligned}$$

Expressing $-\lambda n^2$ from these conditions, one obtains:

$$\begin{aligned}
-\lambda n^2/2 &= \frac{(1-\rho^{I})S_b^2}{c_{12}^{I} + c_{12}^{II}m} + \left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{m(c_{12}^{I} + c_{12}^{II}m)} \\
&= \frac{(1-\rho^{II})\bar{S}_w^2}{m^2 c_{12}^{II}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&(1-\rho^{I})S_b^2 Mm^2 c_{12}^{II} + (M-m)(1-\rho^{II})\bar{S}_w^2 mc_{12}^{II} \\
&\qquad - (1-\rho^{II})M\bar{S}_w^2(c_{12}^{I} + c_{12}^{II}m) = 0, \\
&[(1-\rho^{I})S_b^2 Mc_{12}^{II} - (1-\rho^{II})\bar{S}_w^2 c_{12}^{II}]m^2 \\
&\qquad - [(1-\rho^{II})M\bar{S}_w^2 c_{12}^{I}] = 0, \\
m &= \sqrt{\frac{(1-\rho^{II})M\bar{S}_w^2 c_{12}^{I}}{[(1-\rho^{I})S_b^2 M - (1-\rho^{II})\bar{S}_w^2]c_{12}^{II}}} \\
&= \sqrt{\frac{c_{12}^{I}}{c_{12}^{II}}\frac{(1-\rho^{II})\bar{S}_w^2}{(1-\rho^{I})S_b^2 - (1-\rho^{II})\bar{S}_w^2/M}}.
\end{aligned}$$

From the survey budget (3.6),

$$n = \frac{C_0}{c_{12}^{I} + c_{12}^{II}m} = \frac{C_0}{c_{12}^{I} + \sqrt{\dfrac{(1-\rho^{II})\bar{S}_w^2 c_{12}^{I}c_{12}^{II}}{(1-\rho^{I})S_b^2 - (1-\rho^{II})\bar{S}_w^2/M}}}.$$

Finally, the variance of the difference estimator is

$$
\begin{aligned}
&\mathrm{V}_{e,o}[d] \\
&= \frac{2}{C_0}\left(c_{12}^{\mathrm{I}} + \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}}\right) \\
&\quad \times \left[(1-\rho^{\mathrm{I}})S_b^2 \right. \\
&\qquad \left. + \left(\sqrt{\frac{c_{12}^{\mathrm{II}}}{c_{12}^{\mathrm{I}}}\frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} - \frac{1}{M}\right)(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}}\right. \\
&\qquad + (1-\rho^{\mathrm{II}})\bar{S}_w^2\sqrt{c_{12}^{\mathrm{I}}c_{12}^{\mathrm{II}}\frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} \\
&\qquad + \left[(1-\rho^{\mathrm{I}})S_b^2 - \frac{1}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] \\
&\qquad \times \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}} \\
&\qquad \left. + (1-\rho^{\mathrm{II}})\bar{S}_w^2\left(c_{12}^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M}\right)\right\} \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}.
\end{aligned}
$$

*Proof of Proposition* 7. Ignoring the finite population correcting terms of the order $O(N^{-1})$ and $O(M^{-1})$, equation (3.3) can be written as:

$$
\begin{aligned}
\mathrm{V}_{e,i}[d] &\approx \frac{4\left(c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}}c^{\mathrm{II}}\bar{S}_w^2/S_b^2}\right)}{C_0}\left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}}\bar{S}_w^2 S_b^2\right)\right] \\
&= \frac{4}{C_0}\left(c^{\mathrm{I}}S_b^2 + c^{\mathrm{II}}\bar{S}_w^2 + 2\sqrt{c^{\mathrm{I}}c^{\mathrm{II}}S_b^2\bar{S}_w^2}\right) \\
&= \frac{4}{C_0}\left(\sqrt{c^{\mathrm{I}}S_b^2} + \sqrt{c^{\mathrm{II}}\bar{S}_w^2}\right)^2.
\end{aligned}
$$

Likewise, equation (3.8) can be written as

$$
\begin{aligned}
\mathrm{V}_{e,o}[d] &\approx \frac{2}{C_0}\left[(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}}\right. \\
&\qquad \left. + 2\sqrt{c_{12}^{\mathrm{I}}c_{12}^{\mathrm{II}}(1-\rho^{\mathrm{I}})S_b^2(1-\rho^{\mathrm{II}})\bar{S}_w^2} + (1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{II}}\right] \\
&= \frac{2}{C_0}\left[\sqrt{(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}}} + \sqrt{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{II}}}\right]^2.
\end{aligned}
$$

The statement of Propostion 7 follows immediately from these two expressions.

## References


Binder, D.A., and Hidiroglou, M.A. (1988). Sampling in time. In *Handbook of Statistics*, (Eds., P.R. Krishnaiah and C.R. Rao), North Holland, Amsterdam, 6, 187-211.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed., New York: John Wiley & Sons, Inc.

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

Ernst, L.R. (1999). The maximization and minimization of sample overlap problems: A half century of results. Technical report, U.S. Bureau of Labor Statistics.

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*. New York: John Wiley & Sons, Inc.

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

Kish, L. (1995). *Survey Sampling*, 3rd Ed., New York: John Wiley & Sons, Inc.

Lehtonen, R., and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, *Statistics in Practice*, 2nd Ed., New York: John Wiley & Sons, Inc.

Mas-Colell, A., Whinston, M.D. and Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, UK.

McDonald, T.L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85, 277-292.

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

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

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

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

Scott, C.T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, 8(2), 228-233.

Thompson, M.E. (1997). Theory of Sample Surveys. *Monographs on Statistics and Applied Probability*, New York: Chapman & Hall/CRC, 74.

Thompson, S.K. (1992). *Sampling*, New York: John Wiley & Sons, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd Ed., New York: Springer.