Optimal weighting and overlap for composite estimation in repeated surveys

Mehran, Farhad
Independent Consultant
112 Route de Florissant
Geneva 1206, Switzerland
mehranxfarhad@yahoo.com

Fallah Mohsenkhani, Zohreh
Statistical Research and Training Center
145 Fakouri Street
Tehran 1413-717911, Iran
zohrehf@srtc.ac.ir

Composite estimation has been developed to take advantage of the correlation between common units in repeated surveys with overlap samples, Binder and Hidiroglou (1988). The results presented here differ from conventional methods in a number of ways. First, the correlation is assumed to be known or at least not estimated from the sample data directly. Second, the overlap is considered exact, in the sense that the number of common units is fixed by design not allowing more common units to enter the sample by chance. Third, the framework is set such that it is applicable where there are only two overlap samples, not necessarily a sequence as in the case of rotation sampling schemes, or where the two surveys are not necessarily conducted over time but jointly at the same time for example a labour force survey conducted along a household income and expenditure survey with a fraction of sample units in common.

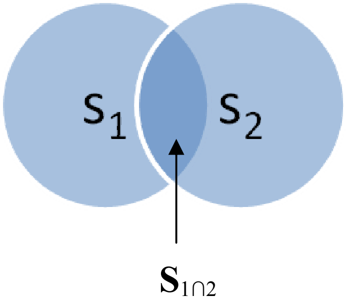
Repeated surveys

Consider a survey on a finite population of N units conducted at two time periods 1 and 2. The population is assumed to be closed during the time interval and the purpose of the survey is to estimate the total value of a characteristic of interest at both time periods. Let *Y* denote the characteristic of interest and t1 and t2 the unknown totals to be estimated:

$$t_1 = Y_{11} + Y_{12} + \cdots + Y_{1N}$$
$$t_2 = Y_{21} + Y_{22} + \cdots + Y_{2N}$$

where Y_{ij} represents the value of *Y* for individual j at time i.

At time 1, a simple random sample of n units is drawn from the population, s_1 and divided randomly into two sub-samples: one, $s_{1\cap 2}$, containing a specific fraction of sample units δ and the other the remaining fraction (1-δ). At time 2, the first sub-sample $s_{1\cap 2}$ is retained to be surveyed again together with a fresh sub-sample drawn from other units so that the full second sample s_2 also consists of a simple random sample of n units from the population with exact sample overlap of δ.



It is important to emphasize that in this framework the new sub-sample in S_2 is drawn from units not already selected in S_1. This formulation is in line with practice in surveys with sample overlap conducted in certain national statistical offices. The traditional framework for composite estimation assumes that the subsample in S_2 is drawn independently from S_1. This means that certain units, other than those in the sample overlap fixed by design, may fall by chance in common in the two samples as they theoretically receive a non-zero probability of selection. For a population of size N, and a sample size of n units and an overlap of m units, the sample space under the present framework would contain $\binom{N}{m}\binom{N-m}{n-m}\binom{N-n}{n-m}$ possible samples against $\binom{N}{m}\binom{N-m}{n-m}\binom{N-m}{n-m}$ under the traditional framework. Although the two frameworks converge for large sample sizes, the differences may be significant in surveys where the sample overlap is implemented at the level of primary sampling units.

Composite estimation of levels

Based on the sample observations, $y_{11}, y_{12}, \ldots, y_{1n}$ for the first sample and $y_{21}, y_{22}, \ldots, y_{2n}$ for the second sample, estimates of the total population values at time 1 and 2 are derived using the conventional method, giving respectively the direct estimates,

$$\hat{t}_1 = \frac{N}{n}\sum_{j \in s1} y_{1j} \text{ and } \hat{t}_2 = \frac{N}{n}\sum_{j \in s2} y_{2j}$$

and their variance,

$$Var(\hat{t}_1) = Var(\hat{t}_2) = N^2\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}$$

where σ^2 denotes the variance of a population unit (Y_{ij}) assumed to be constant over units and time.

If the sample overlap is not void (δ>0), an indirect estimate of the total at time 2 can be derived by adding to the direct estimate at time 1, the estimate of change that occurred between time 1 and 2 using information on the matched sub-sample $s_{1\cap 2}$. The resulting forward estimate is

$$\tilde{t}_2 = \hat{t}_1 + \frac{N}{\delta n}\sum_{j \in s1\cap 2}(y_{2j} - y_{1j})$$

A similar backward estimate for the total at time 1 can also be derived by deducting the estimate of change from the direct estimate at time 2,

$$\tilde{t}_1 = \hat{t}_2 - \frac{N}{\delta n}\sum_{j \in s1\cap 2}(y_{2j} - y_{1j})$$

The backward estimate would serve in situations where the repeated surveys are ad-hoc, not part of a series, or at the start of a sample rotation scheme when no preceding sample exists.

The direct and indirect estimates may be combined to produce a third estimate of the population totals, called composite estimates,

$$t_1^c = (1 - K)\hat{t}_1 + K\tilde{t}_1 \text{ and } t_2^c = (1 - K)\hat{t}_2 + K\tilde{t}_2$$

where (1-K) and K are the weights attached to the direct and indirect estimates, respectively. The composite estimate reduces to the direct estimate for K=0 and to the indirect estimate for K=1.

The weight that minimize the variance of the composite estimates is given by

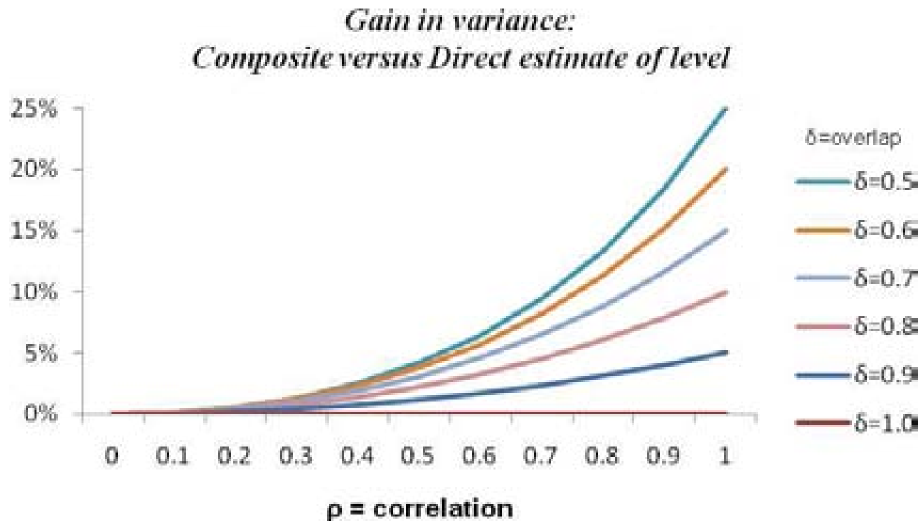
$$K = \frac{\delta\rho/2}{1 - \rho + \delta\rho}$$

where ρ denotes the correlation between the common units at the two time periods 1 and 2. When there is no sample overlap (δ=0) or the correlation is zero (ρ=0), the optimal weight is also zero (K=0), and the composite and the direct estimates are the same. Also, for complete sample overlaps (δ=1), the value of the optimal weight (K = ρ/2) is immaterial as the composite and direct estimates are identical for any value of the weight. Finally, if the correlation is perfect (ρ=1), the composite estimate is the arithmetic average of the direct and indirect estimates.

The variance of the composite estimate of level for t_1 and t_2 are equal to each other and given by

$$Var(t_1^c) = Var(\hat{t}_1) - N^2 \frac{\delta(1-\delta)\rho^2/2}{1-\rho+\delta\rho}\frac{\sigma^2}{n}$$

The relative gain in variance by using the composite estimate instead of the direct estimate when the sampling fraction is negligible is shown graphically below for different degrees of overlap and correlation coefficients,



It can be observed that for full sample overlap (δ=1) or zero correlation between the survey rounds (ρ=0), there is no gain in using the composite over the direct estimate. Maximum gained is achieved when the correlation is one (ρ=1) and there is some overlap between the samples.

For a given non-zero correlation $(\rho \neq 0)$ the optimal overlap that minimizes the variance of the composite estimate with optimal weights is given by

$$\delta = \frac{\sqrt{1-\rho}-(1-\rho)}{\rho}$$

The corresponding relative gain in variance of using the composite estimate with optimal overlap and optimal weights instead of the direct estimate for negligible sampling fraction is $(1-\sqrt{1-\rho})^2/2$.

Composite estimation of change

The change between time periods 1 and 2 can be estimated on the basis of the sample overlap alone

$$\Delta_{12} = \frac{N}{\delta n}\sum_{j\varepsilon s1\cap 2}(y_{2j}-y_{1j})$$

or the change in the direct estimates $\hat{t}_2 - \hat{t}_1$ or the change in the indirect estimates $\tilde{t}_2 - \tilde{t}_1$ or in the composite estimates

$$t_2^c - t_1^c = (1-2K)(\hat{t}_2 - \hat{t}_1) + 2K\Delta_{12}.$$

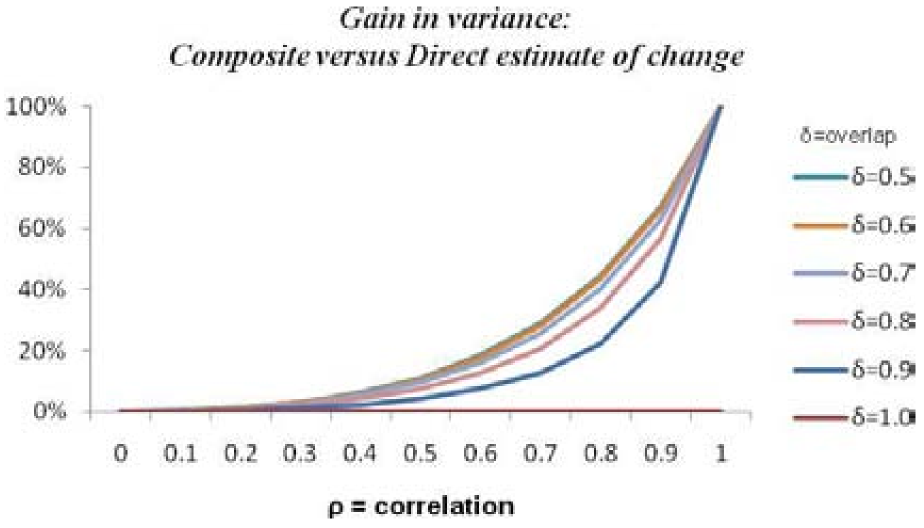
It can be shown that the optimal weight K derived earlier for the composite estimate of level is also the optimal weight for the composite estimate of change. This is an advantage as one of the criticism raised against the traditional framework for composite estimation is the difference between the optimal value of K for estimating level than the optimal value for estimating change.

The variance of the composite estimate of change with optimal weight is given by

$$Var(\Delta_{12}^{c}) = Var(\hat{\Delta}_{12}) - N^2 \frac{2\delta(1-\delta)\rho^2}{1-\rho+\delta\rho}\frac{\sigma^2}{n}$$

where $Var(\hat{\Delta}_{12}) = 2N^2\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n} - 2N^2(\delta - \frac{n}{N})\frac{\rho\sigma^2}{n}$.

The relative gain in variance for different degrees of overlap and correlation coefficients when the sampling fraction is negligible is shown in the following diagram. It can be noted that composite estimation is relatively more efficient for estimation of change than for estimation of level. For the same sample overlap and correlation coefficient, the relative gain in variance is larger for the composite estimate of change than composite estimate of level.



As in the case of estimation of change, for full sample overlap (δ=1) or zero correlation between the survey rounds (ρ=0), there is no gain in using the composite over the direct estimate. The maximum gain of 100% is achieved when the correlation is one (ρ=1) for any degree of overlap (δ<1).

For a given non-zero correlation $(\rho \neq 0)$ the optimal overlap that minimizes the variance of the composite estimate with optimal weights is one-half (δ=½). The corresponding relative gain in variance of using the composite estimate with optimal overlap and optimal weights instead of the direct estimate for negligible sampling fraction is $\rho^2/(2-\rho)^2$.

Applications

- National labour force surveys. Composite estimation is commonly used in national labour force surveys with rotation sampling design. The method presented here has been developed for application to the quarterly labour force survey of Iran, Fallah Mohsenkhani *et al.* (2009). The sample design has a rotation pattern (2-2-2) according to which sample units are in the survey for two consecutive quarters, leave the survey in the next two quarters, and return in the survey for the subsequent two quarters before leaving the survey altogether. Based on the sample overlap between two consecutive quarters (δ=½), composite estimators of the main labour force variables of the current quarter are computed using the corresponding optimal K values (K=0.278 male employed; K=0.216 female employed; K=0.123 male unemployed; K=0.131 female unemployed; K=0.297 male inactive; and K=0.207 female inactive). These composite estimators are then used as benchmark for calibrating the sampling weights of the current quarter. The calibrated weights provide the final weights and are applied to all variables of the current quarter's survey.

Different types of composite estimators have been applied in other national labour force surveys. An example is the composite estimation used in the US Current Population Survey (CPS) to improve the monthly employment and unemployment estimates, US Bureau of Labor Statistics (2006). The sample design has a rotation pattern (4-8-4) that entails 75% common units between two consecutive months. The CPS composite estimator involves two parameters, the composite parameter K and an additional coefficient A, meant to adjust for the bias associated with time in sample, the result of interviewing the same CPS respondents several times. The values of the parameters are fixed and derived empirically, separately for the unemployed (K=0.4, A=0.3) and the employed (K=0.7, A=0.4).

Another example of composite estimation is the regression composite estimator of the Canadian monthly labour force survey, Singh, Kennedy and Wu (2001). The survey follows a rotation scheme with six rotation groups, and in any two consecutive months, five of the rotation groups form the overlapping sample. A key element of the regression composite estimator is the use of micro-matching to relate past and current data on the overlap sample. The method presented here as well as the ak-estimator used in the US CPS is based on the relationship between past and present data at the macro-level.

The Australian monthly Labour Force Survey has recently introduced a more general composite weighting estimator, Australian Bureau of Statistics (2007). The rotation pattern of the survey is designed to maintain in the sample the sampling units for 8 months, with one-eighth of the sample being replaced each month. The composite weighting estimator takes into account the correlation structure between the current month's and the previous six months' data, Bell (2001).

- Global estimate of child labour. The International Labour Office (ILO) has produced global estimates of child labour at several occasions since 1995, the latest of which referred to the developments between 2004 and 2008, Diallo, *et al.* (2010). The 2008 estimate was calculated as a composite estimate based on two samples of countries, one, called the full-sample, consisting of 50 national datasets used to produce a direct estimate of child labour in 2008, and the other, called the matched-sample, consisting of a sub-sample of 27datasets of countries common both in 2004 and 2008 or with repeated surveys during the period. The data were stratified by region and the optimal weights for composite estimation were calculated at the regional level. The relative gain in variance due to composite estimation in 2008 was 26% for Asia and the Pacific, 10% for Latin America and the Caribbean, 8% for Sub-Saharan Africa, and 21% for the rest of the world.

- Joint surveys. Another application of composite estimation as formulated here is for joint surveys with overlap samples. Consider, for example, a labour force survey (LFS) and a household income and expenditure survey (HIES) conducted jointly with a fraction of the sample common in both surveys. Each survey gives an estimate of the number of employed persons, the LFS based on a specialized questionnaire and the HIES based generally on a reduced, simplified sequence of questions. Given the correlation between employment as measured in LFS and in HIES, composite estimation may be used to improve the estimates from each survey as suggested by the following expressions,

$$t^{c}_{LFS-HIES} = (1-K)\hat{t}_{LFS} + K(\tilde{t}_{HIES} - \tilde{t}_{LFS})$$

$$t^{c}_{HIES-LFS} = (1-K)\hat{t}_{HIES} + K(\tilde{t}_{LFS} - \tilde{t}_{HIES})$$

where $\hat{t}_{LFS}$ and $\hat{t}_{HIES}$ are the direct estimates of employment based on the LFS and HIES surveys respectively, and $\tilde{t}_{LFS}$ and $\tilde{t}_{HIES}$ are the corresponding estimates based on the common units alone.

REFERENCES

Australian Bureau of Statistics (2007). Information Paper: Forthcoming Changes to Labour Force Statistics, Australia, Catalogue No. 6292.0.

Bell, Ph. (2001). "Comparison of Alternative Labour Force Survey Estimators," Survey Methodology, Vol. 27, No. 1, pp. 53-63.

Binder, D.A. and Hidiroglou, M.A. (1988). "Sampling in time." Handbook of Statistics, 6: Sampling, Elsevier Science, NY, 187-211.

Diallo, Y., Hagemann, F., Etienne, A., Gurbuzer, Y. and Mehran, F. (2010). Global child labour developments: Measuring trends from 2004 to 2008, International Labour Office, Geneva.

Fallah Mohsenkhani, Z., Harandi, F., Golchi, S., Bidar Bakhtnia, A. and Mehran, F. (2009). "Improved estimates of labour force statistics using the rotation pattern of the labour force survey," Statistical Research and Training Center, Tehran, Iran.

Singh, A.C., Kennedy, B. and Wu, S. (2001). "Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design," Survey Methodology, Vol. 27, No. 1, pp. 33-44.

US Bureau of Labor Statistics (2006). Current Population Survey Design and Methodology, Technical Paper 66, http://www.bls.census.gov/cps/tp/tp66.htm, Chapter 10, p. 11.

RÉSUMÉ

L'estimation composite est utilisée pour améliorer les estimateurs des variables courantes des enquêtes répétées où le plan d'échantillonnage comprend des unités communes entre enquêtes successive. La méthode consiste à combiner deux estimateurs initiaux d'une même variable, un basé sur seules les données de l'enquête courante et l'autre sur une mise-a-jour de l'estimateur précédent avec les données des unités communes des deux enquêtes successive. Cet article donne l'expression des poids optimaux pour combiner les estimateurs initiaux dans le cas d'échantillonnages aléatoires simple de populations finies et fermées. L'article présente également le degré optimal de chevauchement d'échantillons en function de la correlation entre les enquêtes successive, et examine si les valeurs optimales obtenues pour l'estimation de niveau sont aussi optimales pour l'estimation d'évolution. Finalement, certaines applications sont présentées dans le contexte des enquêtes nationales de force de travail avec plans d'échantillonnage avec rotation et l'estimation globale du travail des enfants basée sur des échantillons de pays á deux périodes différentes.