COMBINING DATA FROM DEPENDENT SOURCES: A CASE STUDY OF THE LABOUR FORCE SURVEY AND THE CANADIAN COMMUNITY HEALTH SURVEY

Kate Wilder[1], Steven Thomas[2]

ABSTRACT

Several options exist for combining data collected from multiple surveys. In the case of the Canadian Community Health Survey (CCHS) and the Labour Force Survey (LFS), the samples are selected from the same area frame, with a large overlap of primary sampling units (PSUs), creating dependence. This paper explores the options for estimating a population parameter using two such samples and gives the results of a study in which sample weights were created for a combined sample. Variance estimation is also explored.

KEY WORDS: Bootstrap , Dependent Samples, Integration, Weighting.

RÉSUMÉ

Il existe plusieurs options pour l'intégration des données recueillies d'enquêtes multiples. Dans le cas de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et l'Enquête sur la population active (EPA), les échantillons sont sélectionnés d'une même base aréolaire, avec un chevauchement important pour les « unités primaires d'échantillonnage » (UPE), ce qui crée un effet de dépendance. Cet article explore différentes options pour estimer un paramètre en utilisant deux échantillons dépendants, et présente les résultats d'une étude dans laquelle les poids ont été créés pour l'échantillon combiné. L'estimation de la variance sera aussi discutée.

MOTS CLÉS : Échantillons dépendants; pondération; intégration; bootstrap.

1. INTRODUCTION

1.1 Overview

A new strategy is currently being developed to estimate disability rates by province, age group and gender. Several options for this strategy are envisaged, but the main idea is to add a short set of questions (the Disability Screening Questions or DSQ module) onto existing surveys, in particular the Labour Force Survey (LFS) and the Canadian Community Health Survey (CCHS). It is thought that this could be an easy, inexpensive way to collect information from a large sample in a timely manner. Other surveys have already taken advantage of variations on this option by interviewing LFS respondents during the LFS interview (a live supplement), or by surveying LFS or CCHS respondents after the main interviews have been conducted (a follow-up). The challenge arises when we want to produce estimates of disability rates using the respondents from both surveys. Could a composite estimator be used? What about a pooled approach, where the samples are combined and weighted as one? Is it possible to correctly estimate a joint selection probability? Could the LFS and CCHS respondents be combined to create a sampling frame for a follow-up survey? How should their weights be calculated? The concept of pooling LFS and CCHS respondents to create a frame for second phase surveys is not new. Laflamme and Landry (2009) also investigated this option. Their results are mentioned briefly throughout the paper.

1.2 Some background on each survey

[1] Kate Wilder, Statistics Canada, Household Survey Methods Division, R.H. Coats Bldg., 16th floor, 100 Tunney's Pasture Driveway, RHC-16J, Ottawa ON, K1A 0T6, Kate.Wilder@statcan.gc.ca
[2] Steven Thomas, Statistics Canada, Household Survey Methods Division, R.H. Coats Bldg., 16th floor, 100 Tunney's Pasture Driveway, RHC-16M, Ottawa ON, K1A 0T6, Steven.Thomas@statcan.gc.ca

1.2.1 The Labour Force Survey

The LFS is a monthly survey which is used to estimate labour market indicators such as the unemployment rate. Its sample of 54,000 households is selected from an area frame of clustered dwellings using a stratified, multi-stage design. The 6,600 or so sampled clusters are selected with probability proportional to size from over 1,000 design strata. Within each cluster, a systematic sample of dwellings is selected (called a 'start'). Sampled dwellings remain in sample for six months, with approximately one sixth rotated out and replaced each month. All eligible members (aged 15+) of the household are included in the survey. Response to the LFS is mandated by law.

1.2.2 The Canadian Community Health Survey

The CCHS is a voluntary, cross-sectional survey that collects information related to health status, healthcare use and health determinants for the Canadian population. A new sample of approximately 10,500 households is selected every two months, using the same area frame as the LFS as well as a telephone list frame. From the area frame, a design similar to the LFS is used. The main differences with the sample selection are that clusters are stratified by approximately 120 Health Regions and multiple starts can be selected from the same cluster. From the telephone frame, a stratified simple random sample of landline telephone numbers is selected. For both frames, one member aged 12+ is selected to complete the survey.

1.3 LFS and CCHS Sample Dependence

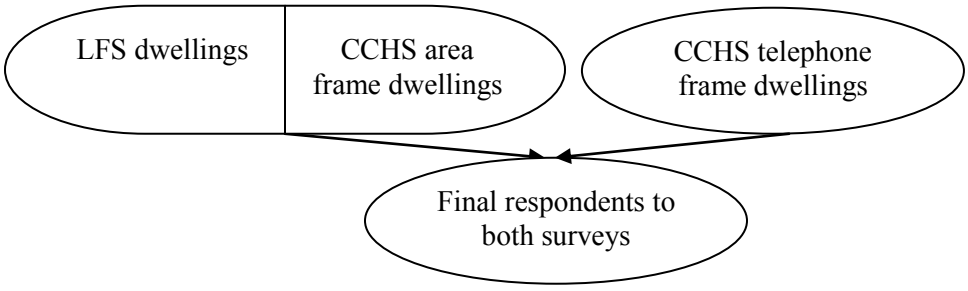
The CCHS selects roughly 40% of its final sample from the area frame. In order to reduce listing costs, the CCHS attempts to select its area frame sample from the same set of clusters as those selected by LFS, causing the two samples to be dependent. The CCHS will select some clusters not sampled by LFS in order to meet the sample size requirements for each health region (HR), but overall there is large overlap. For example, CCHS's sample from November 2009 – October 2010 had 4,257 of 4,746 clusters (90%) in common with the June 2010 sample for the LFS. Laflamme and Landry (2009) have shown that this cluster overlap creates non-negligible covariance between estimates coming from each survey. For this reason, using a simple linear composite estimator to estimate disability rates could be difficult, since the variance of the estimator would also need to be calculated. The bulk of this paper will focus on another approach, termed the "combined-frame" approach, concentrating on issues with the calculation of design weights, nonresponse adjustments, integration of the area and telephone frame samples, and variance estimation.

2. METHODS FOR COMBINING LFS AND CCHS SAMPLES

2.1 The Combined-Frame Approach

Since the area frame sample of the CCHS is coordinated with the LFS sample, one school of thought says that we can combine CCHS and LFS clusters and create initial weights for the dwellings in this combined area frame sample. These dwellings can then be combined with the telephone frame units and multiple-frame adjustments can be applied to obtain a final weight (see Figure 2.1). The main drawback to this method is that it is virtually impossible to calculate a joint CCHS-LFS selection probability due to the differing sample designs and stratification of each survey. The main benefit is that any replication method for variance estimation will take the cluster overlap into account (see section 2.1.4). A feasibility study was undertaken in which 12 months of CCHS sample was combined with one month of LFS sample at the cluster level, in order to determine the main challenges with this method and to estimate the resulting design effects for the combined sample.

Figure 2.1: The combined-frame approach



2.1.1 Initial Weighting

LFS 'basic' weights reflect the probability of selecting a particular group of dwellings (or start) *i*, within a particular cluster *j*, from a particular stratum *h*. Since, within an LFS stratum, the groups of dwellings are the same size and selected from clusters that are selected with probability proportional to size, the LFS sample is initially self-weighting where all dwellings within an LFS stratum have the same basic weight. Since a combined sample contains a different number of clusters per stratum and starts per cluster than the LFS sample alone, the selection probability for each start in the combined sample is modified. We can see this with the following formulae. To calculate an initial weight for the combined sample, an adjustment factor is applied to the basic LFS weight. As noted above, one issue is that the CCHS uses a different stratification than the LFS. For the feasibility study, adjustments were performed at the more detailed LFS stratum level. Although this does not reflect the exact joint probabilities, it should take some of the joint probability into consideration.

For the LFS, the first stage selection probability of cluster *j* in stratum *h* is:

$$\pi_{1hj} = \frac{n_{1h} R^*_{hj}}{\sum_{j \in h} R^*_{hj}} \text{ , where } \quad R^*_{hj} = \left[\text{int}\left(\frac{\tilde{M}_{hj}}{\tilde{m}_h} \right) \right],$$

n_{1h} is the number of clusters selected in stratum *h*, $\tilde{M}_{hj}$ is the expected number of dwellings in cluster *j* in stratum *h*, and $\tilde{m}_h$ is the planned number of dwellings selected per cluster in stratum *h*. An approximation to the first stage selection probability for the combined sample is given by:

$$\pi^*_{1hj} = \frac{n^*_{1h} R^*_{hj}}{\sum_{j \in h} R^*_{hj}},$$

where $n^*_{1h}$is the number of clusters in the combined CCHS/LFS sample in stratum *h*.

For the LFS, the second-stage selection probability of dwelling *i* in stratum *h* is:

$$\pi_{2hji} = \frac{1}{R^*_{hj}}.$$

For the *combined* sample, the second-stage selection probability reflects the increase in the number of starts selected per cluster. Let m^*_{hj} be this new number of starts selected in cluster *j*. The increased second-stage selection probability is:

$$\pi^*_{2hji} = \frac{m^*_{hj}}{R^*_{hj}}.$$

The probability of selecting dwelling *i* in the combined sample is then the product of the first and second-stage probabilities. The inverse of this probability is the combined basic weight:

$$w_i^{Combined} = \frac{\sum_{j \in h} R^*_{hj}}{n^*_{1h} * m^*_{hj}}.$$

In practise, the LFS basic weights are multiplied by the adjustment a_{hji} shown below. A summary of this initial adjustment is given in Table 2.1.1.

$$a_{hji} = \frac{n_{1h}}{n^*_{1h}} \frac{1}{m^*_{hj}}.$$

Table 2.1.1: Clusters and Starts and the Initial Weight Adjustment for the Combined Approach

Survey	# Clusters in sample	# Starts in sample	Mean weight adjustment a_{hji}
LFS (Jun '10)	6 670	6 670	0.62

CCHS (Nov '09- Oct '10)	4 746	5 745	0.44
Overall	**7 159 (4 257 overlap)**	**12 415**	**0.53**

2.1.2 Household Nonresponse Adjustment

After the initial weight adjustment, dwelling weights must be adjusted for household nonresponse, that is to say, the case where no information is obtained from the household. The challenge with the combined sample is that the LFS and CCHS have different response mechanisms due to the nature of the surveys. These differences may be due to the LFS being mandatory while the CCHS is not. Also, nonresponse may be related to the survey topic. To deal with nonresponse for a combined LFS/CCHS sample, there are two main options. The first option is to re-form the adjustment classes using a segmentation method or a more intricate modeling approach. For example, the CCHS method could be adopted. This method uses geographic information and information obtained during the data collection process to model a particular household's response propensity and create weight adjustment classes based on these groups (See Sarafin (2007)). Another option is to keep the existing nonresponse classes from each survey independent. It is unclear which adjustment method is preferable, as more research is needed in this area to better understand the relationship between disability characteristics and nonresponse. Also in practice, some respondents may answer the main LFS or CCHS questions but not the disability screening questions, meaning additional weight adjustments could be necessary.

2.1.3 Integration of Area Frame and Telephone Frame Samples

For the CCHS, during integration of the area frame and telephone frame samples, household weights are adjusted to account for the fact that the both frames cover a large, overlapping portion of the population. In short, the weights of the households selected from this common portion of the area frame (that is, they also have a phone number on the CCHS telephone frame) are adjusted by an integration factor α and the weights of the households selected from the telephone frame are adjusted by the factor $1- \alpha$, $0 \leq \alpha \leq 1$. Weights of the units selected from the "non-common" portion of the area frame are not adjusted. See Wilder (2010) for more details on the integration process.

It is assumed that the current methodology used by the CCHS could also be used on a combined CCHS-LFS sample. One challenge lies in determining which CCHS and LFS area frame units are also covered by the CCHS telephone list frame.

To determine which households (both LFS and CCHS) are found on the CCHS telephone frame, a matching exercise is undertaken, using phone numbers provided at the time of interview. In rare cases (< 10%), when a phone number is not available for a household and the respondent did not state whether they have a listed telephone number, the probability of belonging to the list frame is modeled using logistic regression. One bonus with the LFS sample is that the quality of the phone numbers of the LFS respondents is high; the LFS phone numbers are verified since they are used to contact the respondents month after month. The phone numbers obtained from CCHS respondents are used to link with the telephone frame and are not used for follow-up (Table 2.1.3).

Table 2.1.3: LFS and CCHS Telephone Frame Coverage Rates

Survey	Area frame households that provide tel. number	# of tel. numbers linked to CCHS telephone frame	Overall % linked to CCHS telephone frame
LFS (Jun '10)	51,054* / 53,538 (95%)	36,955 / 51,054 (72%)	69%
CCHS (Nov '09- Oct '10)	26,982 / 32,517 (83%)	20,817 / 26,982 (77%)	64%

*includes cell phone numbers

The second challenge when integrating the area frame households with the telephone frame households is determining a value for the integration factor, α. For CCHS production, α is fixed at 0.4, a value which represents the approximate proportion of the 'listed landline' sampled households which was selected from the area frame. For this study, α was increased to 0.61 to reflect the increase in sample coming from the area frame due to the addition of the LFS respondents. Otherwise, the survey weights would be too variable and a disproportionate amount of weight would be given to the CCHS-selected telephone frame units, leading to higher variance in the sample weights.

2.1.4 Variance Estimation

Currently the LFS uses the jackknife method to estimate the variance. The CCHS uses the bootstrap method, selecting *n-1* of *n* sampled clusters[3] per LFS stratum on the area frame, and *n-1* of *n* sampled phone numbers per CCHS stratum on the telephone frame to replicate potential samples. It was decided that for the combined approach, the CCHS methodology could be used. For the area frame, selecting the clusters for each replicate has the beneficial effect of keeping all selected starts from a cluster either in or out of a replicate. The dependence between LFS and CCHS samples (the fact that multiple starts from the same cluster are selected) is taken into account and variance estimates will better reflect reality. For the units selected from the telephone frame, the usual CCHS methodology was used (*n-1* of *n* telephone numbers per health region are selected with replacement) to create the replicates. All subsequent weight adjustments are either applied to or recalculated for each replicate, mimicking what is done to the main sample. Telephone frame and area frame replicates are integrated using the same integration factor (i.e. 0.61) for each replicate. Household weights are then calibrated to known household counts at the province by household size level (1,2,3+ persons).

2.1.5 Some Combined-Frame Simulation Results

At the time of writing this paper, the Disability Screening Questions module had not been collected through either of the surveys, so there was limited common content with which to produce combined estimates. A few common variables included marital status and household size. For the feasibility study, one person aged 15+ was selected randomly from each household in the LFS sample, which is likely what would be done if we were to proceed with a live supplement option on the LFS. This led to final sample sizes of 53 538 and 58 564 respondents aged 15+ from the LFS and CCHS respectively. Weight adjustments were applied to reflect this selection.

Table 2.1.5: CVs and Design Effects for the Combined-Frame approach

Characteristic	*n* in sample with characteristic	$\hat{t}$	$\hat{p}$	Coeff. of variation (bootstrap)	Design Effect (bootstrap)
Separated/Divorced	23,563	3,476,719	0.13	1.07	1.53
Household size=2	42,836	9,640,125	0.36	0.71	2.35

These estimates of *t* and *p* are consistent with the estimates produced using solely one CCHS or LFS sample.

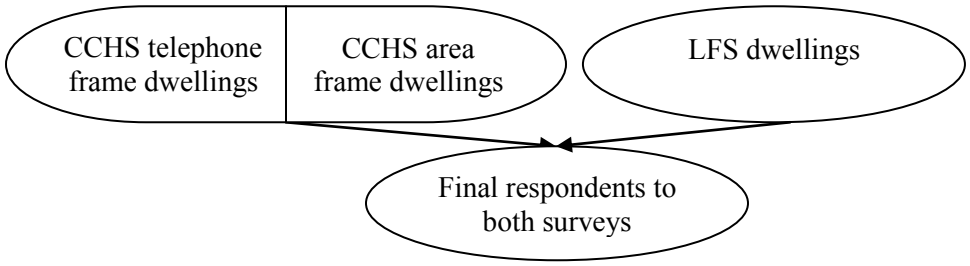
2.2 An Alternative Estimator (Dual-Frame Approach)

As an alternative to the combined frame approach, we could use a composite (dual-frame) estimator to estimate a population parameter using the respondents of the LFS and CCHS. It could take the form:

$$\hat{Y} = \theta\hat{Y}_{LFS} + (1-\theta)\hat{Y}_{CCHS}\ , 0 \le \theta \le 1$$

We can visualize this method of combining the two samples using Figure 3.3:

Figure 3.3: The dual-frame approach



In theory, the disability rates produced using either survey should be similar, although it is expected that the context of the two surveys could have an impact on the responses to the disability module. On the CCHS, for example, respondents will have already been asked many questions about their health, reminding them of their particular ailments. Estimates obtained using this dual-frame approach would fall between $\hat{Y}_{LFS}$ and $\hat{Y}_{CCHS}$. If this method were chosen for estimating disability rates, the formula for the variance of the estimates would be:

3 The actual methodology is somewhat more complicated, and involves the LFS rotation group associated with each cluster.

$$v(\hat{Y}) = \theta^2 v(\hat{Y}_{LFS}) + (1-\theta)^2 v(\hat{Y}_{CCHS}) + 2\theta(1-\theta)cov(\hat{Y}_{LFS}, \hat{Y}_{CCHS}).$$

Of course, by design, the LFS and CCHS area frame samples have a large overlap of clusters, and are thus not independent. Laflamme et al. (2009) showed that the covariance term in the above formula is likely greater than zero. If this dual-frame estimator is to be used with these samples, we must estimate the covariance. It could be argued that the coordinated bootstrap or jackknife method, wherein the same PSU (cluster) is selected to be in each survey for a particular replicate, would be sufficient. An exercise was undertaken wherein person-level weights were calculated separately for an annual CCHS sample and a monthly LFS sample, and then adjusted with an integration factor θ of 0.5. For the LFS sample, bootstrap replicates were coordinated with the area frame replicates selected for CCHS. That is, for a given replicate, the same clusters[4] were selected when possible. As was done to the main sample, final bootstrap weights were adjusted by an integration factor of 0.5. The effect of coordinating the bootstrap replicates on the variance of the estimates can be seen in Table 3.3:

Table 3.3: Coefficients of variation for the composite approach and two variance estimation methods

		Composite approach with coordinated bootstrap		Composite approach – no bootstrap coordination	
Characteristic	$\hat{t}$	Stderr	CV	Stderr	CV
Separated/divorced	3,500,495	37,541	1.07	36,353	1.04
Single	8,172,671	48,717	0.6	45,440	0.56
HHsize=1	3,911,667	39,725	1.02	37,856	0.97
HHsize=2	9,469,391	64,652	0.68	59,785	0.63

*Note estimates of totals have changed slightly from those in Table 2.1.5 due to the fact that the 2010 annual CCHS sample (complete with different calibration groups), and not the Nov 2009-Oct 2010 sample, was used when combining with the LFS.

We see from this table that CVs using the coordinated bootstrap approach are slightly higher than those for the non-coordinated method as expected. We also see that the estimates and CVs are similar to those obtained using the combined-frame approach, suggesting that the composite approach with coordinated bootstrap replication may be a viable option.

4. CONCLUSION

More work is needed before adopting either method of combining LFS and CCHS samples to produce estimates of disability rates or a combined sampling frame. A combined-frame approach is feasible and should capture the covariance between LFS and CCHS estimates, but the initial weights will not perfectly reflect the initial selection probability, creating potential bias. Also, a combined-frame estimator may not be the most efficient. Currently, a module on disabilities is in testing on the CCHS and as a live LFS supplement in order to measure differences in the rates produced, but likely the final version of this module will be added to one survey only, and rates will be produced this way.

REFERENCES

Hartley, H. O. (1962), "Multiple Frame Surveys", *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 203-206.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J.(1998). "Methodology of the Canadian Labour Force Survey". *Statistics Canada*.

Laflamme, G., Landry, S. (2009). "Results From a Pilot Survey to Evaluate the Master Sample Methodology". Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, April, 2009.

Sarafin, C., Thomas, S., and Simard, M. (2007), "A Review of the Weighting Strategy for the Redesigned Canadian Community Health Survey" in *2007 Proceedings of the Survey Methods Section,* Statistical Society of Canada.

Wilder, K., and Thomas, S. (2010). "Dual-frame Weighting and Estimation Challenges for the Canadian Community Health Survey". *2010 Proceedings of Statistics Canada International Methodology Symposium.*

4 The coordination of replicates is obtained using rotation groups, of which a particular cluster can only belong to one.