

## ON CROSS-SECTIONAL ESTIMATION FOR REPEATED PANEL HOUSEHOLD SURVEYS

Takis Merkouris<sup>1</sup>

### ABSTRACT

This paper considers weighting and estimation procedures that combine information from multiple panels of a repeated panel household survey for cross-sectional estimation. The dynamic nature of a repeated panel survey is discussed in relation to estimation of population parameters at any wave of the survey. The setting of a repeated panel survey with overlapping panels is described as a special case of a multiple frame survey, with panels forming a nested time sequence of frames. The paper outlines weighting strategies suitable for various multiple panel survey situations. The proposed weighting schemes involve adjustment of weights from domains of the combined panel sample that represent identical time periods covered by the individual panels. The integration of this type of weight adjustment with other weight adjustments required in cross-sectional estimation for a repeated panel household survey is also discussed.

KEY WORDS: Multiple panels; multiple frames; weighting adjustment; generalized regression.

### RÉSUMÉ

Cet article traite de procédures d'estimation et de pondération qui combinent l'information provenant de panels multiples d'une enquête ménage par panels répétés pour de l'estimation transversale. On traite de la nature dynamique d'un sondage par panels répétés en relation avec l'estimation des paramètres d'une population à n'importe quel niveau d'un sondage. La préparation d'un sondage à panels répétés avec des panels de chevauchement est présentée comme un cas spécial d'un sondage à bases multiples, avec des panels formant une suite de bases de sondage emboîtées. Cet article donne un aperçu des stratégies de pondération applicables à une variété de situations d'un sondage à panels multiples. Le modèle de pondération proposé implique l'ajustement des poids par domaine de l'échantillon de panels combinés qui représentent la période de temps identique du panel individuel. On discute également de l'intégration de ce type d'ajustement pondéré avec d'autres types d'ajustements pondérés requis pour l'estimation transversale.

MOTS CLÉS: Panels multiples; bases de sondage multiples; ajustement pondéré; regression généralisée.

### 1. INTRODUCTION

A panel survey collects the survey data for the same sample elements at different time points, or waves. A repeated panel survey is made up of a series of panel surveys each of a fixed duration. The type of repeated panel household survey considered in this paper consists of overlapping panels, with two or more panels covering part of the same time period. A panel survey, though primarily conducted for longitudinal purposes, may also be used to produce cross-sectional estimates, meaning estimates of population parameters at distinct time points.

The process of obtaining cross-sectional estimates at any wave of a household panel survey after the first presents difficulties arising from the dynamic nature of the panel. Weighting schemes involving

adjustments that deal with dynamic aspects of a single panel, such as attrition, movers and cohabitants, have been discussed in the literature; see Kalton and Brick (1995), and Lavallée (1995). Yet, there seems to be a paucity of work in the literature on weighting and estimation for repeated panel household surveys. This paper considers procedures that combine information from multiple panels of a repeated panel household survey for cross-sectional estimation. The setting of a repeated panel household survey with overlapping panels, illustrated by the example of the Canadian Survey of Labour and Income Dynamics (SLID), is described first in Section 2. Next, the coverage of the population at any given wave by the different panels, as well as issues related to the source and the dynamic nature of the sample are discussed. Also discussed in the same section are the analogies with a multiple frame

---

<sup>1</sup> Takis Merkouris, Statistics Canada, Ottawa, Ontario, K1A 0T6, E-mail: merkpan@statcan.ca

setting, and the use of the combined panels as a representative cross-sectional sample. The weighting and estimation problem in repeated panel household surveys is described in Section 3. Weighting strategies suitable for various survey situations are then outlined. Bias and efficiency aspects of various weighting procedures are discussed. The integration of this type of weight adjustment with other weight adjustments required in cross-sectional estimation from a repeated panel household survey is discussed in Section 4. Concluding remarks on the proposed procedures are presented in Section 5.

## 2. THE GENERAL FRAMEWORK

In a repeated panel household survey a sample of households is selected for each panel from the population of households existing at the start of the panel. All the individuals in the sampled households become panel members to be followed throughout the duration of the panel or until they leave the survey population. At a subsequent time, or wave, the household sample consists of all the households in which panel members reside. The type of repeated panel survey household considered in this paper consists of overlapping panels, with two or more panels covering part of the same time period. A typical example of the type of survey considered here is the Canadian Survey of Labour and Income Dynamics (SLID), with two overlapping panels each of duration of six years. In SLID, each new panel is introduced three years after the introduction of the previous one. The sample for each panel is made up of two rotation groups from the Canadian Labour Force Survey (LFS), which uses a stratified multistage design with an area frame in which the households are the final sampling units. It is assumed in this paper that the source of the panel sample for the household panel surveys in consideration is similar to that of SLID. In particular, it is assumed that the samples of the panels are independent, as a result of the random rotation group scheme commonly used in the design of household surveys. Independence of the panel samples is desirable for maximizing the efficiency of cross-sectional estimators of population parameters. The time lag between panels, however, makes this assumption difficult to meet, as some individuals who move (geographically) in the time between the selection of the different panels can be selected in two different panels. This instance of population dynamics may in the course of time be transposed into an instance of sample dynamics, unique to surveys with multiple panels, as members of one panel may move to another panel.

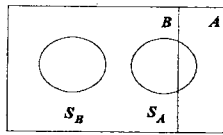
Essential to cross-sectional estimation are changes in the population composition over time, occurring when individuals leave or enter the population. In a single-panel household survey, new entrants who have joined the survey population since the start of the panel, but live in households that do not contain any members of the original population, are not represented in the sample at later waves. A household survey with multiple overlapping panels provides a better coverage of the survey population, since it reduces the time period not covered by any of the panels. In the case of SLID, this reduction is from a maximum of six to a maximum of three years. However, the problem of complete coverage remains, unless a special supplementary sample of the non-covered population is taken. Such a survey procedure, along with the necessary weighting adjustments, is described in Lavallée (1995). A simpler but effective alternative approach involves the selection at any wave of a new sample that covers the entire survey population but does not form a new panel. This kind of supplementary sample (henceforth to be called top-up) is used only once, for cross-sectional purposes, and would normally be of a smaller size. Thus, a household survey with multiple overlapping panels and a top-up sample provides complete coverage of the target population for cross-sectional purposes.

The situation with regard to individuals who leave the population is simple. For any panel, the sampling frame for the survey population at a time point  $t$  is essentially the sampling frame for the population at the start of the panel, with the leavers in the intervening period being treated as blanks on the frame. Panel members, originally selected or cohabitants, who leave the population before time  $t$  correspond to blanks in the frame, and, thus, their effect on cross-sectional estimates at time  $t$  is loss of efficiency but not bias; see also Kalton and Brick (1995). This consideration leads to the following perspective. Although overlapping panels cover part of the same time period, regarding cross-sectional representation each panel covers the entire survey population represented by the preceding panels. Accordingly, the frames of the panels form a nested time sequence, with the frame of each panel containing the frame of the preceding panel. Clearly then, this can be viewed as a special multiple frame survey sampling setting.

The analogy with multiple frame survey sampling places the problem of cross-sectional estimation for repeated surveys with overlapping panels into a familiar framework. There are, however, distinctive characteristics of multiple panel surveys that have to be considered if using the multiple frame

approach to formulate a cross-sectional estimation methodology. In this special case of multiple frame setting, the differences among the frames are induced by time, with the frames of subsequent panels being created over the course of time. The non-overlapping domains of the frames are temporal, and consist of the new entrants in the population between the starts of successive panels. The overlapping domains are identical in size only, since the composition of the population changes within the time between the starts of successive panels. This is in contrast with the static nature of the usual multiple frame survey setting. It should be emphasized, however, that as frame of each panel at any time point is considered the one at the start of the panel, that is, the frame from which the panel was originally selected, but without the leavers. The sample domains are more dynamic. For instance, with the presence of new entrants (originally absent cohabitants) the sample of a panel crosses the boundary of its frame into the frame of the succeeding panel.

For the purpose of developing a cross-sectional procedure that effectively combines information from the panels of a repeated panel household survey, it suffices to consider the simple case of two overlapping panels at the time point of the start of the second panel. Then, using multiple frame notation, with  $B$  and  $A$  denoting the frames of the first and the second panel ( $B \subset A$ ) and with  $s_B, s_A$  denoting the respective samples, the setting can be presented schematically as



In the above diagram, the second panel represents the cross-sectional universe  $U_A$ , so that  $A$  is the complete frame. The overlap domain  $B$  is the remaining of the original frame of the first panel. The domain  $a = B^c \cap A$  consists of all new entrants in the population since the start of the first panel. The samples  $s_B$  and  $s_A$  are the originally selected ones, with  $s_B$  having been reduced because of leavers and non-respondents. The samples  $s_A$  and  $s_B$  are drawn independently from  $A$  and  $B$  according to specified probability designs  $p_A(s_A)$  and  $p_B(s_B)$ , and may be assumed to be disjoint. The terms panel  $A$  and panel  $B$  will be used to denote the two samples at any time point.

### 3. CROSS-SECTIONAL WEIGHTING AND ESTIMATION

This section considers procedures that combine information from multiple panels of a repeated household survey for cross-sectional estimation of population parameters at any wave. The discussion will be confined to estimates of totals. A uniform approach to estimation procedures for households and individuals is presented. The problem of combining information at the estimation stage is essentially the adjustment of the weights from the samples of the separate panels so that a set of proper weights for the combined sample is obtained.

For the construction of a cross-sectionally representative combined sample at any wave, a survey scheme as that of SLID, with a top-up sample taken at each wave, is considered. Then, with the setting as depicted in the diagram, let  $s_{ab} = s_A \cap B$  and  $s_a = s_A \cap a$  denote the sample domains of panel  $A$ . Also, let  $\pi_{Ai}$  and  $\pi_{Bi}$  denote the inclusion probabilities of the  $i$ -th unit (household or any individual within it) for the original samples  $s_A$  and  $s_B$ , respectively. The two samples can be disjoint with respect to selected dwellings, by sampling design, but as noted earlier, individuals (or even households) selected in panel  $B$  who have moved to areas sampled for panel  $A$  can be selected in panel  $A$ . This situation is akin to that of duplicate sample units in multiple frame surveys. In repeated panel household surveys, an operational constraint motivated by respondent burden may be to exclude from  $s_A$  units already selected in  $s_B$ . For a discussion on this, see Lavallée (1994). Here, as in the multiple frame case, it is observed that if the probabilities  $\pi_{Ai}$  and  $\pi_{Bi}$  are small the probability of duplicate units is negligible, and in effect  $s_A \cap s_B = \emptyset$ .

The two sampling designs  $p_A(s_A)$  and  $p_B(s_B)$  induce a well-defined design  $p(s)$  on the set of samples  $s = s_A \cup s_B$  in  $A$ . Then, the two samples can be viewed as selected independently from the frame  $A$  according to the design  $p(s)$ . Thus conventional estimators, based on a single frame, may be constructed from  $p(s)$ . The standard approach is to assign sample units weights made inversely proportional to their inclusion probabilities. The inclusion probability  $\pi_i = P(i \in s)$  of the  $i$ -th unit of the combined sample  $s$  is given by  $\pi_{Ai} + \pi_{Bi}$ , if  $i \in s \cap B$ , and by  $\pi_{Ai}$  if  $i \in s \cap a$ . The weight of the  $i$ -th unit of the sample is then  $w_i = 1/\pi_i$ . Now, let a value  $y_i$  be associated with each population unit  $i$ , and define the population total  $Y_A = \sum_A y_i$ . Then the standard estimator

$$\hat{Y}_A = \sum_s w_i y_i = \sum_{s_B \cup s_{ab}} (\pi_{Ai} + \pi_{Bi})^{-1} y_i + \sum_{s_a} \pi_{Ai}^{-1} y_i \quad (1)$$

of the total  $Y_A$  can be used. An estimator of this form has been proposed by Kalton and Anderson (1986) for multiple frame surveys in which identification of duplicate sample units is not required. Under the assumption of this section regarding duplicate sample units, the estimator  $\hat{Y}$  is approximately equal to the Horvitz-Thompson estimator. The approach just outlined is not in general feasible, since determination of the probability  $\pi_i = \pi_{Ai} + \pi_{Bi}$  for  $i \in s \cap B$  requires that the selection probabilities of the sampled units be known over both frames, which is difficult or impossible to ascertain in household surveys. In multiple panel surveys additional complications arise from the time element. For individuals that move (e.g., to another stratum) in the time between the selection of the panels it is impossible to determine both  $\pi_{Ai}$  and  $\pi_{Bi}$ .

An alternative strategy needs to be considered for developing weights for the overlap sample  $s \cap B$ . An approach that provides a general framework for handling this problem requires information on the probability of inclusion in only one of  $s_A$  or  $s_B$ , thus avoiding the difficulty noted above. The essence of the alternative approach considered here is to associate with the  $i$ -th unit from frame  $B$  a positive constant  $p_i$  if the unit is selected in  $s_B$ , and the constant  $1-p_i$  if the unit is selected in  $s_A$ , and then form the weight of the unit as

$$w_i^* = p_i \frac{1}{\pi_{Bi}} I(i \in s_B) + (1-p_i) \frac{1}{\pi_{Ai}} I(i \in s_{ab}), \quad i \in l \quad (2)$$

where  $I$  is the usual sample membership indicator variable. Clearly,  $E(w_i^*) = 1$ , and thus the use of weights  $w_i^*$  will yield unbiased estimators  $\hat{Y}_B = \sum_B w_i^* y_i$  for the total  $Y_B = \sum_B y_i$  for any choice of positive constants  $p_i$  satisfying  $p_i < 1$ , and for any sampling designs  $p_A(s_A)$  and  $p_B(s_B)$ . Equation (2) can be written alternatively as  $w_i^* = p_i w_{Bi} + (1-p_i) w_{Ai}$ , with the obvious definition of the weights  $w_{Bi}$  and  $w_{Ai}$  associated with the samples  $s_B$  and  $s_A$ . The class of weighting schemes defined by equation (2) consists essentially of different adjustments of the weights of the original samples selected from the frames  $A$  and  $B$ . The intractable single-frame weight  $w_i$  defined as the inverse selection probability  $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$  if  $i \in s \cap B$  can be viewed as a special case of  $w_i^*$  with  $p_i = \pi_{Bi} (\pi_{Ai} + \pi_{Bi})^{-1}$ .

The question arises then as to an alternative, ideally optimal, choice of  $p_i$ , for any  $i \in s \cap B$ . One approach is to choose the  $p_i$  to minimize the variance of the estimated total  $\hat{Y}_A = \sum_B w_i^* y_i + \sum_a w_a y_a$ , where  $w_i = (\pi_{Ai})^{-1} I(i \in s_a)$ . However, minimization of the variance of  $\hat{Y}_A$  with respect to  $p_i$  for  $i \in s \cap B$  is intractable. A simpler option is to restrict the class of weighting schemes defined by equation (2) to one in which the

coefficients are specified not at the unit level but rather at a higher level, which may be a stratum or the complete sample  $s \cap B$ . The case with the same coefficient  $p$  for all units in  $s \cap B$  is considered here. Then, an optimal value of  $p$  can be derived by minimizing the variance of the estimator  $\hat{Y}_A$ , which in view of (2) may take the form

$$\hat{Y}_A = p \hat{Y}_B + (1-p) \hat{Y}_{ab} + \hat{Y}_a, \quad (3)$$

with  $\hat{Y}_B$  and  $\hat{Y}_{ab}$  being independent unbiased estimators of  $Y_B$ . This can be recognized as a special case of a dual frame estimator, in the setting described in the introduction; see Skinner and Rao (1995) for dual frame estimation for complex surveys. Generalization of formula (3) to more than two panels is straightforward. Because of the distinct features of a panel survey, combination of the panels for cross-sectional estimation through a linear combination such as in equation (3) involves a weight adjustment of the original sample units, which is one in a sequence of adjustments leading to the final stage of weight calibration. Weight adjustment schemes based on dual frame methodology, which do not rely on knowledge of the population sizes of the two frames, can be applied in the present context.

An optimal value of  $p$  can be chosen to minimize the variance of  $\hat{Y}_A$ . The optimal value of  $p$  is given in terms of unknown variances and covariances, and needs to be estimated from the sample data. To avoid the dependence of the resulting estimator on the variable of interest  $y$ , the optimization process is applied to some auxiliary count variable, such as the size of frame  $B$ . An alternative approach is based on pseudo maximum likelihood estimation; see Skinner and Rao (1995) for details. These procedures are cumbersome or not applicable in multiple panels.

Another approach is suggested upon writing (3) in the regression form  $\hat{Y}_A = \hat{Y}_{ab} + \hat{Y}_a + p(\hat{Y}_B - \hat{Y}_{ab})$ , where  $p$  can be viewed as a regression coefficient. A generalized regression procedure functioning as a composite weighting adjustment can then be effectively applied to combine data from the two panels. In this version of generalized regression, each panel's weights are adjusted to make composites of the estimates  $\hat{Y}_B$  and  $\hat{Y}_{ab}$  between the panels. The procedure is based on an extension of the constraint system of the generalized regression, in which estimates of comparable totals between two surveys are equated. This form of generalized regression was applied by Zieschang (1990) to the US Consumer Expenditure Survey, and proposed by Singh and Wu (1996) for multiple frame surveys. This method is operationally convenient and

can be readily extended to more than two panels. Elaboration on an adaptation of this method to the present context is beyond the scope of this paper. A notable alternative involves the special choice with  $p=1$ , which yields a simple unbiased but inefficient estimator, since the sample  $s_{ab}$  is not utilized.

It has been assumed thus far that the units of the overlap sample domain  $s_a$  can be identified. The information needed to determine whether the units in  $s_a$  are new entrants to the population, after the start of the previous panel  $B$ , may not be available under the operating procedures of a repeated panel survey. In this situation the weighting process combines the two distinct samples  $s_B$  and  $s_A$  without distinguishing between the domains  $s_{ab}$  and  $s_a$  of  $s_A$ . The estimator then takes the form  $\hat{Y}_A = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_A}$ , with a slight change to a more convenient notation. The effect of this is the under-representation of the new population in the domain  $B \cap A$ , by a factor of  $1-p$ . A calibration of the weights to population totals of the frame  $A$  will correct this with respect to certain auxiliary population characteristics, but some bias may be incurred since the members of the new population (newborns, immigrants) will most likely have different survey characteristics. The size of the domain  $s_a$  is nevertheless very small, relative to  $s_A$ , and the possibility of bias may not be a serious concern, especially when the panel  $A$  is a top-up sample. The optimal value of  $p$  is given now as the ratio of variances  $V(\hat{Y}_{s_A})/[V(\hat{Y}_{s_A})+V(\hat{Y}_{s_B})]$ . A reasonable approximation of the optimal  $p$  can be derived noting that the size of the frame  $B$  may be only a little smaller than the size of the frame  $A$ , and assuming that the variances of the populations covered by these frames are nearly the same. Then, disregarding finite population corrections, it can be shown that a nearly optimal value of  $p$  is given by

$$p = (n_B/d_B)(n_B/d_B + n_A/d_A)^{-1}, \quad (4)$$

where  $n_B$ ,  $n_A$  are the sample sizes of  $s_B$  and  $s_A$ , and  $d_B$ ,  $d_A$  are the design effects associated with the samples  $s_B$  and  $s_A$ . The determination of the optimal value  $p$  requires estimation of the two design effects, which need not be based on the samples  $s_B$  and  $s_A$ . Approximate values of  $d_B$  and  $d_A$  may be available from past surveys. The dependence, however, of the optimal  $p$  on the variable  $y$ , through  $d_B$  and  $d_A$ , requires a compromise solution. To this end, approximate values of  $d_B$  and  $d_A$  could be used for a count variable, such as population size or any count variable associated with a large portion of the survey population and strongly correlated with main survey variables. When the design effects are identical

( $d_A = d_B$ ) the optimal  $p$  is simply  $p = n_B/(n_B + n_A)^{-1}$ , and is independent of any variable of interest. Note that taking the size of  $B$  equal to the size of  $A$  in the approximation of  $p$  gives more weight to the sample  $s_A$ . This may be advantageous considering the possibility of bias noted above, and also because bias due to sample attrition is likely to be larger for the older panel  $B$ . The practical approach just outlined would work well in a situation involving one panel and a top-up sample, or two panels with small time lag in between and a top-up sample, as in SLID. Considering the operational convenience and the parameter-free determination of  $p$  in this approach, its use could be contemplated in these situations even when the overlap sample domains can be identified. In more general situations the appropriate procedures outlined earlier can be applied with minor modifications.

#### 4. OTHER WEIGHT ADJUSTMENTS

The combination of multiple panels for cross-sectional purposes at any wave has been presented thus far as an adjustment of the weights of the originally sampled units in the separate panels that are respondents at that wave. Further weight adjustments are necessary because of the changes in the panels after their first wave. In addition, the usual weight adjustment by which the weights are calibrated so that they sum to known population totals for key demographic characteristics is also performed. The integration of the various weight adjustments is briefly outlined below, in the order imposed by the dynamic nature of the multiple panel survey.

The first adjustment, applied in relation to the original units, is for wave non-response, which arises when a sampled unit responds for some but not all of the waves for which it was eligible. For a discussion on weight adjustment for wave non-response see Kalton and Brick (1995). The adjustment is applied separately on the different panels at each wave.

The second adjustment is for the combination of the samples of the various panels into one sample for cross-sectional estimation. It applies to the weights of the sampled units of the panels, adjusted for wave non-response, and may involve one of the methods outlined in the previous section. Note that this adjustment is specific to each wave because of wave non-response, and involves the original samples  $s_B$  and  $s_A$ .

The third weight adjustment is a more general weighting method, termed weight share method (Lavallée, 1995), whereby at any wave weights are assigned to cohabitants, and to non-sampled households formed after the first wave by members of originally

selected households. The case of selected individuals who have moved to other selected households can also be handled by this weighting method. For details on the weight share method see Lavallée (1995), and Kalton and Brick (1995). It should be noted that in multiple panel surveys the phenomenon of individuals moving from one panel to another panel between waves may be encountered. Thus, the panels are truly distinct only with respect to their first wave.

In the final weight calibration adjustment, sample estimates for demographic characteristics are controlled to independent totals at the time of the current wave, which in the simple case as in the diagram correspond to totals of the frame  $A$ . Then the weights of the combined sample  $s_B \cup s_A$  are calibrated to totals of the frame  $A$ . When the sample domain  $s_a$  can be identified, it is possible to contemplate calibrating the weights of the sample  $s_B \cup s_{ab}$  to totals of the frame  $B$ , and the weights of the sample  $s_a$  to totals of the frame domain  $a = B \cap A$ . Since the size of  $s_a$  is very small (for panels with a small time lag in between) and the number of controls is relatively large, as usual for household surveys, this option is not preferred for efficiency reasons.

## 5. CONCLUDING REMARKS

The weighting procedures outlined in this paper can be used to combine information from multiple panels of a repeated household survey for cross-sectional estimation in a fairly general setting involving panels with a given design. Design issues regarding determination of optimum sampling fractions for the panels, in conjunction with efficient combination of the panel data, have not been considered. These issues merit particular consideration. Given the complexity of a repeated panel household survey, a reasonable combination of efficient estimation procedure and operational convenience should be employed in any particular situation. An empirical study of the comparative merits of alternative weighting schemes

would be helpful in that respect. Finally, it is noted that the quality of a cross-sectional estimation procedure depends on the identification of various overlap sample domains and on design features of the survey, such as the time lag between panels or the use of a top-up sample.

## ACKNOWLEDGEMENTS

The author is grateful to Milorad Kovacevic and Johanne Tremblay for helpful discussions.

## REFERENCES

- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society Ser.A*, 149,65-82.
- Kalton, G., and Brick, J.M.(1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21,33-44.
- Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- Lavallée, P.(1994). Ajout du second panel à l'EDTR: sélection et pondération. Internal document. Statistics Canada.
- Singh, A.C., and Wu, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 69-77.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Zieschang, K. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.