Journal of Official Statistics, Vol. 25, No. 1, 2009, pp. 63–76

On the Inter-Regional Mover Problem in Panel Household Surveys

Takis Merkouris[1]

When a panel household survey is used for cross-sectional analysis, certain weighting and variance calculation issues may arise because of inter-regional moves of selected individuals during the lifetime of the panel. This article points out these difficulties and deals with them in conjunction with the application of the weight share method, which assigns a proper cross-sectional sampling weight to movers and cohabitants. In particular, the weight share method applied separately to each region is explored as an effective means for handling the inter-regional mover problem. The comparative merits of this alternative approach are assessed under various panel household survey designs, and suggestions are made for a suitable approach in each case.

Key words: Cross-sectional estimation; cohabitants; longitudinal household; geographic stratification; weight share method.

1. Introduction

When a panel survey is used to produce cross-sectional estimates of population parameters at distinct time points, called survey waves, certain difficulties related to sample weighting and variance calculations may arise because of moves of sample units during the lifetime of the panel. Such difficulties do not arise in ordinary cross-sectional surveys, but in panel surveys – primarily designed for longitudinal purposes – moving sample units are traced according to specified rules and, thus, may be optionally included in the cross-sectional component of the survey.

Panel household surveys typically employ geographic stratification of the survey population at the time of selection of the sample. The level of geographic stratification that is of interest is the lowest level at which cross-sectional estimates are produced. Such geographic strata may be large administrative regions of a country, such as the American states or the Canadian provinces or areas of the NUTS II territorial classification in the European Union. Adjustments of sampling weights, such as non-response adjustment and calibration, are customarily carried out separately for each of these strata. In the following, the term "regions" will be used to refer to such geographic strata. At subsequent waves, the two prime problems relating to inter-regional moves of sample units are (i) possible large difference in size of

[1] Department of Statistics, Athens University of Economics and Business, Patision 76, Athens 10434, Greece. Email: merkouris@aueb.gr
Acknowledgment: This research was carried out at Statistics Canada. The author is grateful to the Editor and the referees for their constructive comments.

sampling weights among regions, and (ii) sample dependencies among regions, induced by moved units. It should be noted that similar difficulties are encountered in ordinary cross-sectional establishment surveys with outdated frames (see, for example, Rivest 1999); such frames cause sample units to appear in different strata from those from which they were selected. In contrast, in panel surveys the two problematic aspects are intrinsic features of the sample changes over time. In particular, in a panel household survey these two features relate to changes in the composition of the sample households.

A cross-sectional weighting method that deals with changes in the composition of selected households over the lifetime of a panel has been described in Kalton and Brick (1995) and Lavallée (1995). This method, termed *the weight share method*, is a weighting procedure that assigns a basic weight to every individual who at any wave after the first is found in a longitudinal household, that is, a household containing at least one individual selected in the original panel sample. In particular, the weight share method assigns a proper weight to originally selected individuals who have subsequently moved to other longitudinal households.

This article addresses the potential problems resulting from inter-regional moves of selected individuals in a panel household survey in conjuction with the application of the weight share method. To this end, the alternative approach of applying the weight share procedure separately to each region is explored. The characteristic feature of this alternative approach is that it assigns a sampling weight of zero to all individuals in the panel, including the selected movers, who at a subsequent survey wave reside in a region in which they were not present at the time of sampling. The effect of applying the weight share method by region on statistical properties of derived estimators, as well as the operational implications of this procedure, is examined in comparison with the standard weight share method.

For a comprehensive treatment of the inter-regional mover problem, it suffices to consider the simple case of a single-panel household survey, possibly supplemented with a "top-up" sample at some or all later survey waves. A top-up sample means here a supplementary sample that covers the entire survey population at the time of sampling but does not form a new panel. This sample is to be used only once, in combination with the panel, to maintain cross-sectional representativeness for the particular survey wave, and its size would normally be smaller than the panel's size. The more elaborate rotating panel household surveys, such as the Canadian Survey of Labour and Income Dynamics (SLID) and the European Survey on Statistics on Income and Living Conditions (EU-SILC), can be dealt with in a similar manner.

The article is organized as follows. A description of the potential problems resulting from inter-regional moves is given first in Section 2. A general formulation of the weight share method for a single panel is presented in Section 3. The weight share method as applied by region is described in Section 4. A formal comparison of the two weight share procedures is given in Section 5. A discussion of their relative merits, supported by empirical evidence based on data from the SLID, is given in Section 6. Finally, the case of a repeated panel survey with overlapping panels is considered in the concluding Section 7, along with measures for handling the inter-regional mover problem at the design stage of various types of panel survey.

2. Problems Resulting from Inter-Regional Moves

In this section, an account of the potential problems resulting from the inter-regional moves of sampled individuals is given in the order of their relative importance. A more formal description of these problems will be given in subsequent sections.

2.1. Large Differences in Sampling Weights Among Regions

The size of the sampling weights of some inter-regional movers may be very different from the size of a typical sampling weight in their new region. This happens when small regions are oversampled for the purpose of producing regional estimates of sufficient accuracy. For example, in the case of the SLID, in the largest province of Ontario, Canada, there are weights as large as 1,900, while in the much smaller province of Prince Edward Island the weights range from 55 to 72. Problems this may cause relate to confidentiality, nonresponse adjustment of sampling weights, imputation, and to the possibility of erroneous calibrated weights (especially in small poststrata that contain such movers). Furthermore, large differences among weights of movers and weights of original individuals in a region may result in inflated sampling variances for the region, especially for small subpopulations containing movers. In particular, extremely large weights can give rise to very influential sampling units for quantitative data, resulting in unrealistic estimates of high variance.

This problem may necessitate an adjustment of the sampling weights of all movers from the same region into the particular region in which they reside at the present survey wave according to the corresponding demographic count of movers. Such an *ad hoc* adjustment of the sampling weights of inter-regional movers can alleviate these problems but cannot eliminate them, and it may even introduce some bias. It should be noted that accurate and timely demographic information on the total accumulated number of inter-regional movers since the selection of the panel may not be available. Furthermore, given the large number of inter-regional move patterns (region of origin and region of destination – 90 permutations in the case of the SLID), there may be considerable operational complexity associated with this type of adjustment. General procedures of trimming extreme sampling weights (e.g., Potter 1990; Liu et al. 2003) also carry the risk of introducing bias in the estimates and do not seem to be suitable in the present context.

2.2. Sample Dependencies Among Regions

At any survey wave after the first, the variance of a national-level estimator cannot be obtained as the sum of the variances of the regional-level estimators, as may be readily done in purely cross-sectional surveys, because of the covariance terms induced by movers from one region to another.

We can deal with this complication by carrying out variance estimation at the national level, treating the movers as still residing in their original region. Variances at the regional level can then be obtained by treating the current population of any region as domain cutting across the original populations of the various regions. This procedure can be carried out straightforwardly, but at increased operational complexity, with any of the resampling methods (e.g., jackknife or bootstrap) usually employed for variance estimation in household surveys. If the sampling weights are to be calibrated to regional population totals, the weights of the inter-regional movers should be calibrated to population totals of their current region. When using a resampling method of variance estimation, calibration has to be carried out simultaneously for all regions for every replicate in the resampling procedure in order to account for covariances due to inter-regional moves. This may increase the computational time considerably.

The aforementioned problem of variance estimation arises at any level of geographic aggregation. As a general rule, moves among geographic strata should be ignored for the purpose of variance estimation, and variances at any stratum level could be estimated, if needed, by treating these strata as domains.

2.3. Random Size of Regional Sample

The inter-regional movement of sampling units during the lifetime of the panel induces randomness in the realized sample size of each region at each survey wave after the first. This results in some increase in the variability of regional-level estimators, the regions having in effect turned from strata into domains. This effect of the randomness of the regional sample size increases with the number of inter-regional movers.

3. A General Formulation of the Weight Share Method

The following formulation of the weight share method for a single-panel survey will lay the ground for developing an alternative weighting procedure to deal with the problems described in the previous section.

At any survey wave after the first, the weight share method assigns a weight to non-selected individuals who have joined longitudinal households. Following Lavallée (1995), such individuals found in longitudinal households are termed cohabitants. The cohabitants are distinguished into originally present cohabitants, if they belong to the original (sampled) population, and originally absent cohabitants, if they are new entrants to the population. Other problematic situations that can be handled by the weight share method include the weighting for households formed after the first wave by members of different originally selected households, and the weighting of originally selected individuals who have subsequently moved to other longitudinal households. A general formulation of the weight share method is as follows.

Definition 1 (*Weight share method*). Let there be N individuals in the population U_t at a survey wave (time t) after the selection of the panel, with N_i individuals in household $\mathcal{H}_i$ $(i = 1, \ldots, H)$ and $\sum N_i = N$. Let M_i denote the number of individuals in household $\mathcal{H}_i$ at time t who belong to the original population U, with $\sum M_i = M$ denoting the size of the remaining original survey population. One, but not both, of the numbers M_i and $N_i - M_i$ may be zero for any particular household. The sampling weights for the M_i individuals of the original population are defined as random variables that take the value of the inverse of the inclusion probability if the individuals are included in the original panel sample s, drawn with a sampling design $p(s)$, and the value of zero otherwise. On the other hand, the weights for the $N_i - M_i$ individuals not in the original population are identically equal to zero. Formally, the initial sampling weight, w_{i_k}, of the k-th member of household

$\mathcal{H}_i$ is defined as

$$w_{i_k} = \begin{cases} \dfrac{1}{\pi_{i_k}} I(k \in s) & \text{if } k \in U \\ 0, & \text{if } k \notin U \end{cases}$$

where I is the sample membership indicator variable, and π_{i_k} is the probability that the k-th member of household $\mathcal{H}_i$ is included in the sample. Then a common weight for any individual in $\mathcal{H}_i$ at time t is defined as the average of the sampling weights of the M_i members of $\mathcal{H}_i$, that is,

$$w_i = \frac{1}{M_i} \sum_{k=1}^{M_i} w_{i_k} = \begin{cases} \dfrac{1}{M_i} \sum_{k=1}^{M_i} \dfrac{1}{\pi_{i_k}} I(k \in s), & \text{if } M_i > 0 \\ 0, & \text{if } M_i = 0 \end{cases} \quad (1)$$

Now, for a survey variable y, the total for the population of individuals at time t can be expressed as

$$Y = \sum_{i=1}^{H} \sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{H} \sum_{k=1}^{M_i} y_{ik} + \sum_{i=1}^{H} \sum_{k=1}^{N_i - M_i} y_{ik} = Y_o + Y_e \quad (2)$$

where y_{ik} is the value of y for the individual k in household $\mathcal{H}_i$. The two components Y_o and Y_e represent the total for the remaining original survey population and the total for the population of new entrants, respectively. Then, using the weights defined in (1), an estimator of Y, denoted by $\hat{Y}^{NWS}$, is given by

$$\hat{Y}^{NWS} = \sum_{i=1}^{H} w_i \sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{H} w_i \sum_{k=1}^{M_i} y_{ik} + \sum_{i=1}^{H} w_i \sum_{k=1}^{N_i - M_i} y_{ik}$$

The superscript in $\hat{Y}^{NWS}$ indicating the application of the weight share method to the sample at national level is used to distinguish this estimator from an alternative estimator to be introduced in the next section. Note that households composed solely of new entrants (i.e., with $M_i = 0$) are not represented in $\hat{Y}^{NWS}$. On the other hand, it follows from (1) that, for expectation under p(s), $E_{p(s)}(w_i) = 1$ for each i for which $M_i > 0$, and thus

$$E_{p(s)}(\hat{Y}^{NWS}) = \sum_{i=1}^{H} \sum_{k=1}^{M_i} y_{ik} + \sum_{i=1}^{H} \sum_{\substack{k=1 \\ M_i > 0}}^{N_i - M_i} y_{ik} = Y_o + Y_{e_{1+}} \quad (3)$$

where $Y_{e_{1+}}$ denotes the total for the population of new entrants living in households that contain at least one member of the original population. Thus, unbiased estimators for both Y_o and $Y_{e_{1+}}$ are obtained, provided that any originally absent cohabitants can be identified for the correct specification of M_i.

4. The Weight Share Method by Region (RWS)

Seeking to address the problems described in Section 2, we redefine the concepts of mover and cohabitant to be relative to their current region rather than to the entire population.

This, in turn, entails applying the weight share separately for each region. We consider first the decomposition

$$Y = \sum_{r=1}^{R} Y_r = \sum_{r=1}^{R} \left(\sum_{i=1}^{H_r} \sum_{k=1}^{M_{ri}} y_{ik} + \sum_{i=1}^{H_r} \sum_{k=1}^{N_{ri}-M_{ri}} y_{ik} \right)$$

where Y_r is the total for the variable y in region $\mathcal{R}_r$, H_r is the number of households in $\mathcal{R}_r$ at time t, M_{ri} is the number of the original individuals from $\mathcal{R}_r$ who are members of the household $\mathcal{H}_{ri}$ ($i = 1, \ldots, H_r$) at time t, and $N_{ri} - M_{ri}$ is the number of new entrants into $\mathcal{R}_r$ (including movers from other regions) who are members of the household $\mathcal{H}_{ri}$ at time t.

Applying the weight share procedure by region entails treating all $N_{ri} - M_{ri}$ individuals as originally absent from $\mathcal{R}_r$ (not having been selected in $\mathcal{R}_r$), and thus setting their initial sampling weights equal to zero, even for selected movers from other regions. Accordingly, we define a common weight for any individual in household $\mathcal{H}_{ri}$ in $\mathcal{R}_r$ as the average of the sampling weights of the M_{ri} members of $\mathcal{H}_{ri}$, that is,

$$w_{ri} = \frac{1}{M_{ri}} \sum_{k=1}^{M_{ri}} w_{i_k} = \begin{cases} \frac{1}{M_{ri}} \sum_{k=1}^{M_{ri}} \frac{1}{\pi_{i_k}} I(k \in s_r), & \text{if } M_{ri} > 0 \\ 0, & \text{if } M_{ri} = 0 \end{cases}$$

where s_r is the sample from region $\mathcal{R}_r$ drawn with sampling design $p(s_r)$ independently from other regions. Then, an estimator, $\hat{Y}_r^{RWS}$, of the regional total Y_r is given by

$$\hat{Y}_r^{RWS} = \sum_{i=1}^{H_r} w_{ri} \sum_{k=1}^{M_{ri}} y_{ik} + \sum_{i=1}^{H_r} w_{ri} \sum_{k=1}^{N_{ri}-M_{ri}} y_{ik} \tag{4}$$

and an estimator of the national total Y is given by $\hat{Y}^{RWS} = \sum_{r=1}^{R} \hat{Y}_r^{RWS}$.

Households in $\mathcal{R}_r$ with $M_{ri} = 0$ at time t are not represented in $\hat{Y}_r^{RWS}$. Such households may include individuals who are new entrants to the whole population, or movers from other regions, or individuals from both categories. Let then Y_{ro} and $Y_{re_{1+}}$ denote, respectively, the total for the remaining original population in $\mathcal{R}_r$ at time t, and the total for the population of new entrants (including movers from other regions) into $\mathcal{R}_r$ living in households of $\mathcal{R}_r$ at time t that contain at least one member of the original population of $\mathcal{R}_r$.

Proposition 1 The estimator $\hat{Y}_r^{RWS}$ is unbiased for $Y_{ro} + Y_{re_{1+}}$, and the estimator $\hat{Y}^{RWS}$ is unbiased for

$$\sum_{r=1}^{R} (Y_{ro} + Y_{re_{1+}})$$

Proof. It follows from the definition of the weights w_{ri} that $E_{p(s_r)}(w_{ri}) = 1$, for each $\mathcal{H}_{ri}$ in $\mathcal{R}_r$ for which $M_{ri} > 0$. Then

$$E_{p(s_r)}\left(\hat{Y}_r^{RWS}\right) = \sum_{i=1}^{H_r} \sum_{k=1}^{M_{ri}} y_{ik} + \sum_{i=1}^{H_r} \sum_{k=1}^{N_{ri}-M_{ri}} y_{ik} = Y_{ro} + Y_{re_{1+}}$$

Also,

$$E_{p(s)}(\hat{Y}^{RWS}) = \sum_{r=1}^{R} E_{p(s_r)}\left(\hat{Y}_r^{RWS}\right)$$

where $p(s) = \prod_{r=1}^{R} p(s_r)$.□

Since in this alternative weighting procedure the initial sampling weights of movers into $\mathcal{R}_r$ are set equal to zero, the problem described in Section 2.1 does not arise. Also, it follows from (4) and the definition of the weights w_{ri} that the variance of $\hat{Y}^{RWS}$ is obtained additively using the variances of the estimators $\hat{Y}_r^{RWS}$ that is

$$V(\hat{Y}^{RWS}) = \sum_{r=1}^{R} V\left(\hat{Y}_r^{RWS}\right)$$

and thus the variance calculation difficulty described in Section 2.2 is avoided.

The weighting procedure described above is compared with the standard weight share method in detail in the next two sections.

5. Formal Comparison of the Two Weight Share Procedures

For a formal comparison of the weight share procedure applied to the whole (national) sample with the weight share procedure applied by region, rewrite first the weight in (1) for any individual in household $\mathcal{H}_{ri}$ in region $\mathcal{R}_r$ as

$$w_i = \frac{1}{M_i}\sum_{k=1}^{M_i} w_{i_k}$$

$$= \begin{cases} \frac{1}{M_i}\sum_{k=1}^{M_{ri}} \frac{1}{\pi_{i_k}} I(k \in s_r) + \frac{1}{M_i}\sum_{k=1}^{M_i - M_{ri}} \frac{1}{\pi_{i_k}} I(k \in s \backslash s_r), & \text{if } M_i > 0 \\ 0, & \text{if } M_i = 0 \end{cases} \quad (5)$$

where $M_i = \sum_{r=1}^{R} M_{ri}$, and $M_i - M_{ri}$ refers to the number of movers from the other regions. Now, the difference between the two weight share procedures may be clearly seen when estimated regional-level totals for the two procedures are written in terms of the initial sampling weights and household-level totals as follows.

(a) *Weight share for the national sample(NWS)*

In view of (5), the estimated total for y in region $\mathcal{R}_r$ may take the form

$$\hat{Y}_r^{NWS} = \sum_{i=1}^{H_r} w_i y_{ri} = \sum_{i=1}^{H_r}\sum_{k=1}^{M_{ri}} \frac{1}{\pi_{i_k}} I(k \in s_r)\frac{y_{ri}}{M_i} + \sum_{i=1}^{H_r}\sum_{k=1}^{M_i - M_{ri}} \frac{1}{\pi_{i_k}} I(k \in s \backslash s_r)\frac{y_{ri}}{M_i} \quad (6)$$

where $y_{ri} = \sum_{k}^{M_{ri}} y_{ik} + \sum_{k}^{N_{ri} - M_{ri}} y_{ik}$, and $M_i > 0$ for household $\mathcal{H}_{ri}$ in $\mathcal{R}_r$.

Note that the weights in the second term of the summation in (6) may be very different from those in the first term, causing the problems described in Section 2.1. Also, as is evident from (6), there is nonzero covariance between the last term of $\hat{Y}_r^{NWS}$ and the first term of the estimator for a region of origin containing movers to $\mathcal{R}_r$. Clearly, unlike the variance of $\hat{Y}^{RWS}$, the variance of $\hat{Y}^{NWS} = \sum_{r}^{R} \hat{Y}_r^{NWS}$ cannot be obtained additively. Note also that a portion of the variability of $\hat{Y}_r^{NWS}$ is due to the random sample size $\sum_{i=1}^{H_r}\sum_{k=1}^{M_i-M_{ri}} I(k \in s \backslash s_r)$ for the subpopulation of movers (see remark in Section 2.3).

(b) *Weight share by region (RWS)*

It follows from (4) that the estimated total for y in region $\mathcal{R}_r$ may take the form

$$\hat{Y}_r^{RWS} = \sum_{i=1}^{H_r} w_{ri} y_{ri} = \sum_{i=1}^{H_r}\sum_{k=1}^{M_{ri}} \frac{1}{\pi_{i_k}} I(k \in s_r) \frac{y_{ri}}{M_{ri}}$$

where y_{ri} is as in (a), and $M_{ri} > 0$ for household $\mathcal{H}_{ri}$ in $\mathcal{R}_r$. In contrast, $\hat{Y}_r^{NWS}$ in Equation (6) contains an additional term due to movers from other regions. On the other hand, for each household $\mathcal{H}_{ri}$ in $\mathcal{R}_r$ the term associated with the sample s_r is smaller in $\hat{Y}_r^{NWS}$ than the corresponding term in $\hat{Y}_r^{RWS}$ by the factor M_{ri}/M_i. Notice also that terms associated with households for which $M_{ri} = 0$ in $\mathcal{R}_r$ will be missing from the estimator $\hat{Y}_r^{NWS}$.

The two procedures differ, essentially in the construction of the household weights. Explicitly, the weight defined by the NWS procedure for members of household $\mathcal{H}_{ri}$ in region $\mathcal{R}_r$ can be expressed as

$$w_i = c_r w_{ri} + \sum_{\substack{r'=1 \\ r' \neq r}}^{R} c_{r'} w_{r'i}$$

where $c_r = M_{ri}/M_i$, $c_{r'} = M_{r'i}/M_i$, w_{ri} is the average RWS weight of the M_{ri} members of household $\mathcal{H}_{ri}$, and $w_{r'i}$ is the average RWS weight of the $M_{r'i}$ movers from region $\mathcal{R}_{r'}$ in the same household. Prior to the application of the RWS procedure, a weight of zero is assigned to individuals who at time t reside in a region other than the one in which they originally resided. In effect, the RWS procedure treats these individuals in their new region of residence at time t as originally absent. In particular, movers (selected or non-selected in their original region) who are found in longitudinal households in their new region at time t are treated as originally absent cohabitants. On the other hand, the NWS procedure retains the original weights of the selected movers and treats cohabitants coming from another region as originally present.

Consider next the following partition of the population in the region $\mathcal{R}_r$ at time t, denoted by U_t^r,

$$U_t^r = U_0^r \cup \left(U_{e_0}^r \cup U_{e_{1+}}^r\right) \cup \left(U_{m_0}^r \cup U_{m_{1+}}^r\right)$$

where U_0^r denotes the remaining original population in $\mathcal{R}_r$ at time t, $U_{e_{1+}}^r$ denotes the population of new entrants (e.g., immigrants) into $\mathcal{R}_r$ living in households of $\mathcal{R}_r$ at time t that contain at least one member of the entire original population (i.e., original members from $\mathcal{R}_r$ and movers into $\mathcal{R}_r$), $U_{e_0}^r$ denotes the population of new entrants living in households with no member of the entire original population, $U_{m_{1+}}^r$ denotes the population of movers into $\mathcal{R}_r$ living in households of $\mathcal{R}_r$ at time t that contain at least one member of the original population of $\mathcal{R}_r$ or of the population of new entrants into $\mathcal{R}_r$, and $U_{m_0}^r$ denotes the rest of movers into $\mathcal{R}_r$. Now, it is clear from (6) that $E_{p(s)}(w_i) = 1$ for each i for which

$M_i > 0$, and thus the NWS procedure represents, without bias, all the components of U_t^r except the subpopulation $U_{e_0}^r$ (for which $M_i = 0$). Of course, at national level, $\hat{Y}^{NWS} = \sum_r^R \hat{Y}_r^{NWS}$ is the unbiased estimator of the total $Y_o + Y_{e_{1+}}$ defined in (3). On the other hand, by its construction the RWS procedure cannot represent the subpopulations $U_{e_0}^r$ and $U_{m_0}^r$. Moreover, there may be households that contain individuals from both $U_{e_{1+}}^r$ and $U_{m_{1+}}^r$. Therefore, the RWS procedure cannot represent the subset of $U_{e_{1+}}^r$ consisting of new entrants into the population that are cohabitants only of inter-regional movers.

6. The Relative Merits of the Two Approaches

It follows from the discussion in the last two sections that the potential problems described in Section 2 can be addressed using the RWS procedure. On the other hand, the RWS procedure may cause problems of its own. An account of the comparative merits of the NWS and RWS procedures follows.

6.1. Coverage and Bias Considerations

As shown in Sections 4 and 5, in a single-panel survey the RWS procedure cannot represent the population of inter-regional movers who at time *t* live in households that contain no members of the original population of the movers' new region. In fact, the RWS procedure discards the selected movers of this type. When using the RWS procedure, the rest of the inter-regional movers are represented in the panel only by joining households containing at least one selected individual from the original population, whereas when using the NWS procedure these inter-regional movers are sampled in their original region through the use of the frame at the time of the selection of the panel. Clearly, the hit rate for this type of inter-regional movers is lower in the RWS procedure.

The type of movers who are not covered by the RWS procedure usually constitute a relatively very small subpopulation within each region. Over the lifetime of a panel of long duration, however, this subpopulation may become sizeable in some regions. As an illustration, the accumulated number of these movers was estimated over a three-year period in the ten provinces of Canada using available data from the fourth panel of the SLID and cross-sectionally weighted based on the standard weight share method. (At any time the sample of the SLID is made up of two overlapping panels, each one being introduced every three years and used for six consecutive years; see last section for the implications of this with respect to the application of the RWS procedure.) Table 1 shows, at provincial and national level and for three years after the introduction of the panel in 2001, the sample count of these movers as a proportion of the total sample and the weighted sample count as a proportion of the population. An increase in the net number of movers into the various provinces over time is observed. Three years after the introduction of the fourth panel the percentage of the noncovered movers ranged from 0.55 in Quebec (QUE) to 3.45 in Alberta (ALB), while it amounted to 1.40 of the national population. Calibration of the survey weights of the reduced sample (without these movers) to known population totals can lessen any bias effect of this type of noncoverage with respect to characteristics correlated with the calibration variables.

Table 1. Undercoverage (%) of movers by the RWS procedure. SLID data

Province	Year 2002		Year 2003		Year 2004	
	Sample count	Weighted count	Sample count	Weighted count	Sample count	Weighted count
NFL	1.00	1.05	1.17	1.30	0.92	1.08
PEI	0.58	0.66	1.21	1.53	1.89	2.03
NS	0.60	0.50	1.62	1.70	2.35	2.80
NB	0.25	0.19	0.77	0.81	1.30	1.27
QUE	0.16	0.09	0.47	0.33	0.57	0.55
ON	0.69	0.37	1.01	0.60	1.40	0.84
MAN	0.04	0.03	0.41	0.60	0.95	1.31
SASK	0.71	1.21	1.04	1.24	1.45	2.32
ALB	2.29	1.68	4.15	2.38	5.93	3.45
BC	0.64	0.45	1.79	1.52	3.01	2.61
CANADA	0.65	0.47	1.24	0.91	1.81	1.40

It is important to emphasize here that with a top-up sample at any survey wave the subpopulation $U^r_{m_0}$ in each region is covered by the RWS procedure; for the combination of a panel and a top-up sample for cross-sectional estimation, see Merkouris (2001). Estimation for this subpopulation is then possible using the RWS procedure. In particular, any inter-regional movers (selected or nonselected in their original region) found in longitudinal households in their new region at time t will be treated by the RWS procedure as originally present cohabitants, and thus they will be assigned nonzero weights.

As pointed out in Section 5, cohabitants of discarded inter-regional movers are also discarded by the RWS procedure. This is because both these types of household members are assigned initial sampling weights equal to zero. No bias results from this with respect to the originally present cohabitants of the discarded movers in each region, but some risk of bias is associated with the originally absent cohabitants (i.e., immigrants) of the discarded movers, since their subpopulation is not represented in the panel. This subpopulation (the subset of $U^r_{e_{1+}}$ mentioned at the end of Section 5) must be very small, most likely with no representation in the sample, and therefore the potential bias should be negligible. Use of a top-up sample precludes such bias.

In an empirical study of the differences between the NWS and RWS procedures, estimates were produced by each of the two procedures for several characteristics using data from the third wave of the first panel of the SLID. In general, the differences were very small for most characteristics in most regions (provinces). Table 2 shows the percent relative differences (with respect to the NWS procedure) for the few characteristics for which large differences were observed, namely the count of individuals with income below the low income cut-off point (LICO), the total person income, the total income for families of size one, and the corresponding average incomes. For each characteristic the minimum and maximum relative differences by province are shown, as well as the relative difference at national level. The relative differences in the other provinces are closer to the minimum than to the maximum shown in Table 2. The largest differences were associated with characteristics for which the estimated proportion of inter-provincial movers into a particular province was much higher than the estimated overall proportion of

Table 2. NWS vs RWS: Relative differences in estimates (%). SLID data

Variable	Min	Prov	Max	Prov	Canada
LICO	0.16	PEI	6.12	NB	1.35
Income	0.01	SASK	1.41	ALB	−0.03
Income Fs1	−0.06	MAN	8.44	ALB	1.50
Av.Income	0.02	QUE	1.35	ALB	−0.04
Av.Income Fs1	0.03	MAN	6.50	ALB	0.25

inter-provincial movers into that province. For example, the proportion of movers into New Brunswick (NB) who had income below the LICO was more than three times larger than the overall proportion of movers into that province. The corresponding relative difference for the number of individuals with income below the LICO was 6.1%. The potential for bias is, of course, larger for these characteristics. However, the few observed large relative differences do not necessarily indicate bias of the same scale. They may be explained to a large degree by the sampling variability associated with inter-provincial movers whose sampling weight is of very different size from that of a typical weight in their new province, and to a lesser degree by the random size of the corresponding sample; see relevant discussion in Section 2. In this empirical study, the bias effect could not be separately estimated since the true values of the estimated parameters were not known and the estimates were based on a single sample. It should be noted here that estimates for the small subpopulations not covered by the RWS are not likely to be a survey requirement. The issue then is the effect on estimates for larger subpopulations, such as provinces or the entire nation. At the national level the observed relative differences were very small (see last column in Table 2).

6.2. *Variance Considerations*

Regarding the efficiency of regional-level estimators, an analytical assessment of the relative efficiencies of the two procedures is generally intractable for the part of the cross-sectional population that is covered by both procedures. It is fair, though, to say that the two procedures may not differ appreciably because of the very small number of households that contain at least one member from the original population of a region and movers from another region. When a top-up sample is used, and the entire cross-sectional population is thus covered by both procedures, the RWS procedure may result in some loss of efficiency due to discarding selected inter-regional movers living in households with no member from the original population of the movers' new region. This efficiency loss may become noticeable for some regions over the lifetime of the panel, depending on the duration of the panel. As shown in Table 1, in the fourth panel of the SLID the accumulated number of these movers over a three-year period represents 1.81% of the total panel size. The percentage of these movers by province ranges from 0.57 in Quebec to 5.93 in Alberta.

On the other hand, in view of the discussion in Section 2.1, the NWS procedure may result in appreciable loss of efficiency if the differences between the weights of any type of inter-regional movers and the weights of individuals in the new region of the movers are large. And in any case, as indicated in Section 2.3, the NWS procedure makes a rather inefficient use of the selected movers discarded by the RWS procedure due to the random size of the corresponding part of the sample.

In the study based on the SLID data, differences in coefficients of variation (CV's) were calculated for an empirical comparison of the two weight share procedures in terms of efficiency of derived estimators. Because the RWS procedure applies to a slightly smaller population, an assessment of the relative efficiencies based on CV's is more appropriate than the assessment based on variances. The difference in CV's was negligible in all provinces for most study characteristics. Table 3 shows the absolute difference in CV's for the same survey characteristics for which the relative differences in the estimates produced by the two procedures were large. For each of these characteristics, the numbers in Table 3 represent the maximum loss or maximum gain in terms of CV's, by province, associated with the use of the RWS procedure. The maximum loss is very small in all reported cases. It is interesting that for almost all characteristics (LICO being the exception) for which relative differences in estimates were large in Table 2, and in the same provinces, the RWS procedure resulted in an efficiency gain despite the relatively large proportion of discarded units under the RWS procedure in these cases. This seems to support the aforementioned arguments regarding the effect of the weight differential among regions and the inefficient use of the selected movers by the NWS procedure.

6.3. Operational Considerations

Regarding the operational requirements, the RWS procedure can be carried out in a straightforward manner. It only requires knowledge of whether a cohabitant comes from another region in order to classify this cohabitant as originally absent. This classification is not an issue when a top-up sample is used and the RWS procedure is applied after the combination of the top-up sample with the panel, for then all cohabitants are originally present; see Merkouris (2001). However, unbiased estimation requires that inter-regional movers in the top-up sample be identified, as new entrants into their new region, so as to be excluded from the weight adjustment needed for combining the top-up sample with the panel. The identification of the inter-regional movers in the top-up sample may not be possible under the particular operational procedures of a panel household survey. In that case, the nonidentifiable subpopulation of movers will be slightly undercovered. Such an identifiability problem does not arise in the NWS procedure.

On the other hand, if the NWS procedure is to be applied, the sampling weights of inter-regional movers may have to be adjusted if the size of these weights is very different from

Table 3. NWS vs RWS: Absolute Differences in CV's. SLID data

Variable	Max Loss	Prov	Max Gain	Prov
LICO	0.61	NB		
Income	0.07	BC	0.44	ALB
Income Fs1	0.53	PEI	3.00	ALB
Av.Income	0.09	PEI	0.45	ALB
Av.Income Fs1	0.56	PEI	3.98	ALB

the size of a typical sampling weight in the movers' new region. As mentioned in Section 2.1, this adjustment, which should be made before the weight share procedure, may entail considerable operational complexity.

7. Discussion and Summary

Certain complexities associated with inter-regional movers arise in the cross-sectional use of panel household surveys. An alternative weighting approach that involves applying the weight share method separately to each region eliminates these problems, but it may cause other problems because it involves discarding a certain type of inter-regional movers from the cross-sectional estimation process. The relative merits of the two weighting approaches presented above can be summarized in terms of operational convenience, bias, and variance as follows. The alternative approach is considerably more efficient operationally, though it requires the identification of inter-regional movers in a top-up sample when such a supplementary sample is used. It may incur a very small bias, because of undercoverage of the population of movers, or no bias at all if a top-up sample is used. Its effect on statistical efficiency may be a negligible loss, or even some gain for quantitative variables if the differential in sampling weights among regions is large.

The approach to take in a particular panel survey with regard to the inter-regional mover problem depends on relevant features of the survey. Thus, when estimates are required only at national level, the main problem associated with the standard approach is the more complex calculation of variances (see Section 2.2). Of course, this holds even if the survey employs a self-weighting (or nearly self-weighting) sampling design; if regional estimators were also required, some increase in their variability would be an additional concern (see Section 2.3). In a single-panel survey without a top-up sample, the duration of the panel may be an important factor in choosing the most advantageous approach because since the number of movers tends to accumulate over time. Thus, in a panel of long duration, the impact of the inter-regional moves on the quality of estimates produced by either weighting procedure may be substantial, and the trade-off between bias and variance may be more difficult to gauge. The relative accuracy of these estimates could be assessed on the basis of their mean squared error (MSE), if some estimate of the bias component of the MSE is workable. On the other hand, use of a top-up sample ensures the unbiasedness of estimators produced by the alternative RWS approach. Of course, operational considerations (see Section 6.3) should be taken into account in the choice of procedure.

Repeated panel surveys using a series of overlapping panels of fixed duration with a rotational design have been introduced in recent years. Prime examples are the Canadian SLID and the European EU-SILC. For a description of the SLID and the EU-SILC, see Lavigne and Michaud (1998) and Eurostat (2001), respectively. Cross-sectional estimation in multiple-panel household surveys is discussed in Merkouris (2001). The RWS procedure can be adapted to such panel surveys in a straightforward manner. Because of the time overlap of the panels, the population of movers not covered by the RWS procedure is reduced. In the case of the SLID, the duration of each of its two component panels is six years, but the maximum time that may not be covered by a panel is two years. It is important to point out that in multiple-panel surveys in which the introduction of each new panel coincides with each survey wave, as in the EU-SILC, the application of the RWS procedure would have no bias effect on estimators. The latter survey scheme generalizes the sampling structure of a single-panel survey supplemented with a top-up sample at each survey wave.

The above considerations suggest that the problems of inter-regional moves may be preferably best dealt with at the design stage of a panel survey, to the extent allowed by the survey's substantive objectives. Thus, a large differential in sampling weights within the same region caused by inter-regional movers would be avoided with the use of a self-weighting scheme or by constraining the differential in selection probabilities in the various regions. In addition, choosing a broad geographic stratification with a small number of strata would be effective in limiting the number of inter-regional moves. The unbiasedness of the RWS procedure when the introduction of each of a number of overlapping panels coincides with each survey wave could be taken into account when considering options for tracing rules and for the set of individuals to use in cross-sectional analysis. For instance, for considerations of timeliness, one option for the EU-SILC (see Eurostat 2001, p. 76) does not include movers in the cross-sectional weighting. It should be stressed, however, that an effective use of the RWS procedure requires that the survey design include the necessary specifications regarding identification of inter-regional movers in a top-up or in each new panel in a rotating panel survey, if such survey design schemes are used.

8. References

Eurostat (2001). Statistics on Income and Living Conditions (SILC), Sampling. Document EU-SILC 52/01. Eurostat, Luxemburg.

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.

Lavigne, M. and Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Working Paper SLID 98-05 E, Statistics Canada.

Liu, B., Ferraro, E.W., and Brick, J.M. (2003). Trimming Extreme Weights in Household Surveys. Proceedings of the American Statistical Association, Section on Survey Research Methods, 3905–3912.

Merkouris, T. (2001). Cross-sectional Estimation in Multiple-panel Household Surveys. Survey Methodology, 27, 171–181.

Potter, F. (1990). A Study of Procedures to Identify and Trim Extreme Sampling Weights. Proceedings of the American Statistical Association, Section on Survey Research Methods, 225–230.

Rivest, L.P. (1999). Stratum-Jumpers: Can We Avoid Them? Proceedings of the American Statistical Association, Section on Survey Research Methods, 64–72.

Received May 2004
Revised August 2008