THE METHODOLOGY OF THE WORKPLACE AND EMPLOYEE SURVEY

Z. Patak, M. Hidiroglou and P. Lavallée
Zdenek Patak, 11-F, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

Key words: Linked samples, multi-stage design, survey process efficiency.

1. INTRODUCTION AND OBJECTIVES

In recent years Statistics Canada has steadily increased its capacity to follow businesses and individuals longitudinally. This is being achieved by linking survey and administrative data longitudinally to explore the impact of a rapidly changing labour market on firms and their employees. Clients interested in the competitive position of Canadian industry have sponsored surveys on technology use, innovation and the success of small and medium sized enterprises. These surveys are beginning to shed some light on firm growth and decline, particularly how the adaptive and innovative capacities of firms contribute to their success.

Canadian firms and their employees have always faced a competitive, changing environment. The development of a North American free trade zone has certainly heightened awareness of the competitive environment. The growing disparity among workers, in terms of both earnings and hours, has been well documented. These trends contribute to a general sense that economic change is increasingly difficult to understand, that the cost of change is mainly borne by less adaptable workers, and that even among the winners in the labour market, employment is becoming less stable. Looking at these trends, analysts in Statistics Canada and elsewhere have reached the conclusion that there are two key elements missing in our understanding of firm performance and worker outcomes.

The determinants of how well firms respond to change can be properly studied in a longitudinal setting that covers all the firm characteristics and behaviours related to performance. The practices and policies related to employees are also of interest, since they must be the agents of change in the firm. Their fortunes are tied to what they do on the job and how they interact with the internal forces of change within a firm. Thus the ideal survey instrument would follow the linked samples of employers and employees over an indefinite period.

The Workplace and Employee Survey (WES) is such an instrument. The objective of WES is to investigate the relationships among competitiveness, innovation, technology use and human resource management on the employer side, and technology use, training, job stability and earnings on the employee side. WES will shed some light on what triggers hiring and separations, actual and perceived job stability, which employees use particular technologies and how it affects their skill requirements and pay, and how employee compensation and human resource practices relate to firm performance.

In Section 2 we discuss the longitudinal strategy for the employer and employee portions of WES. The sample design aspects are detailed in Section 3. Collection, edit, and imputation strategies are described in Section 4. We provide some detail on weighting and estimation in Section 5, followed by a discussion of future work in Section 6.

2. LONGITUDINAL STRATEGY

The Workplace and Employee Survey has been designed to be a longitudinal survey. It uses two distinct sampling units, workplace and employee. The workplace is defined as a physical location where employer-employee data can be linked directly. These units will be surveyed on a number of successive occasions. The longitudinal strategy for each unit is discussed next.

2.1. Workplace Portion

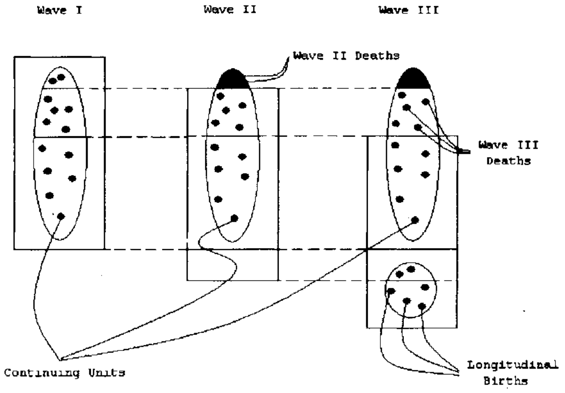
The first production wave of WES is a cross-sectional sample of Canadian workplaces and their employees. The collection of data from this sample will start in the spring of 1999. Subsequent waves will be carried out at one-year intervals. The focus in the first year will be on cross-sectional analysis. To incorporate a time component into the first wave, we will consider post-stratifying the workplace sample by the number of years a unit has been in operation. In the absence of respondent supplied historical data, this will allow us to conduct a limited, time dependent analysis in the absence of multiple waves of data.

The employer portion of WES is carried out in a series of two-wave cycles. For each cycle, the first wave is a typical cross-sectional survey. Only complete and partial respondents are contacted for the second wave. First wave non-respondents are not contacted until wave three. This cycle is repeated every odd survey occasion, i.e., years three, five, seven, and so on.

Starting with the second wave, the emphasis will shift from cross-sectional to longitudinal analysis. All live longitudinal units will be carried forward from wave one to wave two and thereafter. Total non-response units will not be contacted during the second wave data collection. For wave three, every effort will be made to convert refusals. This strategy coincides with our approach to sampling employer births and redrawing of employees in year three, and is explained below.

Prior to the third wave, the panel of continuing units will be expanded by a sample of births that have accrued since the first wave. This sample augmentation will occur before each odd survey occasion. The selected birth units are considered longitudinal. With their addition we will achieve cross-sectional representativity at the start of each two-year cycle. Workplaces with zero employees (eg., owner-operators) are out of scope for WES. If these units acquire employees sometimes in the future, then they will be considered births and sampled accordingly.

Figure 1. Longitudinal sampling of workplaces



The workplace sample has been designed to be robust longitudinally. A small study looking at birth and death rates for Canadian businesses has revealed that the expected sample attrition is approximately 20% a year. To reduce the possibility of a non-certainty stratum being completely eroded over time, the minimum sample size has been set to ten. If, despite these precautions, sample attrition substantially depletes a particular stratum, then the possibility exists that a second panel will be selected to add units within that stratum.

Selected workplaces will remain in sample for a period of at least four years. Their stratum affiliation will be kept fixed until the completion of the fourth wave. We anticipate that during this time the response rates will drop due to response fatigue, and stratification will deteriorate as a result of changes in the business universe. These phenomena are not confined to WES alone; they are experienced by all longitudinal surveys.

The response burden on small employers and their employees will be monitored with each new wave. Currently there are no plans to introduce rotation of small units. This strategy will be reconsidered if response fatigue causes significant erosion in some strata.

To deal with obsolete classification of workplaces we will look at the impact of re-stratifying their population and redrawing the sample after the completion of the fourth wave. We would maximise the overlap between the two samples to keep as many of the existing workplaces in the new sample as possible. If stratum membership changes substantially over time then even maximising the overlap between two successive samples may not be enough to keep all existing units without boosting the sampling fractions. Calibration methods would be considered if a significant number of longitudinal units were to be lost.

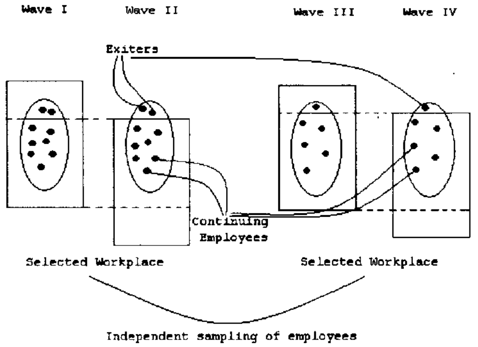
2.2. Employee Portion

Interviewers draw a sample of three, six, nine, or twelve employees at the selected workplaces depending on the size of the workplace. This sampling is limited to every other year because it is a major expense for the survey. In the second year employees who have not changed workplaces will receive the same questionnaire. The rest will be administered an exit questionnaire.

A small but significant portion of workplaces does not keep a list of new hires. Rather than implementing two different sampling strategies depending on what employee information we can get from the workplace, we opted to completely redraw the employee sample in the third year. The two samples will overlap to a certain degree depending on the number of employees in the workplace. For example, the overlap will be 100% for workplaces with three or fewer employees.

The employees reselected in the new sample will provide employee variable estimates for the original longitudinal population for at least four years or until they leave their workplace or the workplace closes down.

Figure 2. Longitudinal sampling of employees.



3. SAMPLE DESIGN

WES has been implemented as a stratified two-stage design to satisfy the survey objectives. We select a sample of workplaces at the first stage and a sample of employees from the selected workplaces at the second stage. The details of the design are discussed next.

3.1. Workplace Portion

The target population for the workplace component is defined as all workplaces operating in Canada with paid employees, with the following exceptions:

a) workplaces located in Yukon or North West Territories;

b) workplaces operating in the following industries: agriculture and related industries; fishing and trapping; highway, street and bridge maintenance; government services; private households; religious organisations

This population is partitioned into non-overlapping groups based on the workplace's industrial activity and geographic region. The workplaces in each industry/region group are allocated to three size strata based on employment. This stratification uses a model-based approach (Godfrey, Roshwalb and Wright, 1984).

With this approach it is assumed that we can use a model defined by $y_i = \beta x_i + \varepsilon_i$, where $\varepsilon_i \approx (0, \sigma^2 x_i^\gamma)$, and that the variable of interest y (eg., wages and salaries) has a variance proportional to a known auxiliary variable x (eg., employment). Several choices of γ are possible. For the ratio estimator we use $\gamma = 1$.

We want to minimise the design variance of the resulting estimator to arrive at a nearly optimal stratification by size. Within each industry/region group we arrange the population units in an ascending order of their auxiliary variable. Then we divide each industry/region subpopulation into H groups of units such that their sums of the auxiliary variables raised to the power γ are approximately equal; that is, $\sum_{i \in U_{g1}} x_{(i)}^\gamma \cong \cdots \cong \sum_{i \in U_{gH}} x_{(i)}^\gamma \cong \sum_{U_h} x_i^\gamma / H$.

The sample is then allocated to the strata by the *equal parts rule*, that is, $n_h = n / H$ in each stratum. For model-based stratification, values other than one can be used for γ . Mahalanobis (1952) stated that $\gamma = 2$ worked well for many skewed distributions. For WES, we tried several values of γ ranging from 1.0 to 2.5.

The classical model-based stratification and allocation procedure was modified for WES. The *equal parts rule* for allocation was dropped in favour of Neyman allocation. This resulted in a more efficient sample.

Once the population is stratified by size, stratum sample sizes are computed using Neyman allocation subject to the constraint that expected marginal design coefficients of variation (CV) not exceed a given CV at the region by industry level. For WES this CV was initially set to 0.10. The resulting allocation yielded a sample of 6,281 workplaces with expected industry-level CVs ranging from 0.01 to 0.09, and expected region-level CVs ranging from 0.02 to 0.04.

The sample size was still well within the budgeted 7,500 units. This allowed us to lower further the expected marginal CVs at the region by industry level to 0.09 by boosting the sample size from 6,281 to 7,378 workplaces. This resulted in lowering expected industry CVs to between 0.036 and 0.040, and expected region CVs to between 0.046 and 0.061. A sample of workplaces was then drawn independently in each stratum using simple random sampling without replacement. Several strata containing very large workplaces were sampled exhaustively producing some 292 certainty units.

Several other sample stratification and allocation methods were investigated. Amongst them were the cumulative $\sqrt{f}$ rule followed by Neyman allocation, and the Lavallée-Hidiroglou method (Lavallée and Hidiroglou, 1988). Below we compare the modified model-based stratification and allocation scheme based on $\gamma = 2$ with the cumulative $\sqrt{f}$ rule, and the Lavallée-Hidiroglou method using Neyman and square root allocations.

To make the comparison fairer in terms of the suggested number of certainty units to the optimal Lavallée-Hidiroglou method, the cumulative $\sqrt{f}$ rule and modified model-based probability strata with weights less than or equal to 1.25 were converted to certainty strata. This was done in anticipation of perhaps introducing sample rotation sometimes in the future. Using a time-out constraint of one year, a workplace with a weight of 1.25 would be rotated out for one survey occasion after being in sample for four years. Lower weights yield longer in sample periods. It was felt that the reduction in response burden due to rotating a unit out for one year every fifth survey occasion and not rotating it out at all was negligible.

Table 1. Comparison of allocation schemes.

Region/ Scheme	LHN	MMB_2	$AMMB_2$	CS	ACS
1	918	930	940	1,385	1,398
2	1,365	1,579	1,595	2,410	2,421
3	1,424	1,649	1,667	2,993	3,067
4	887	910	919	1,363	1,371
5	977	1,090	1,131	2,673	2,833
6	1,097	1,220	1,222	2,066	2,118
Total	**6,668**	**7,378**	**7,474**	**12,890**	**13,208**
Take-all	2,682	292	793	0	1,883

Going from left to right the sample stratification and allocation strategies shown in Table 1 are: (1) Lavallée-Hidiroglou using Neyman allocation (LHN); (2) modified model-based approach using $\gamma = 2$ (MMB_2); (3) cumulative $\sqrt{f}$ rule followed by Neyman allocation (CS); (4) adjusted modified model-based approach using $\gamma = 2$ ($AMMB_2$); (5) adjusted cumulative $\sqrt{f}$ rule followed by Neyman allocation (ACS). The adjustment to $AMMB_2$ and ACS is as described in the preceding paragraph.

The table clearly shows that the CS and ACS are quite inefficient. The sample sizes required to satisfy the marginal CV constraints are nearly double the Lavallée-Hidiroglou (LH) method, which is the optimal procedure. The MMB_2 and $AMMB_2$ fair much better increasing the optimal sample by only ten percent. Where MMB_2 and $AMMB_2$ have a clear advantage over the LH approach is in the area of certainty units. The LH procedure requires that nearly 40% of the sample size be selected with a probability equal to unity. This can place an undue response burden on many medium-to-large units.

Figure 3. Stratum boundaries for model-based and Lavallée-Hidiroglou stratification.

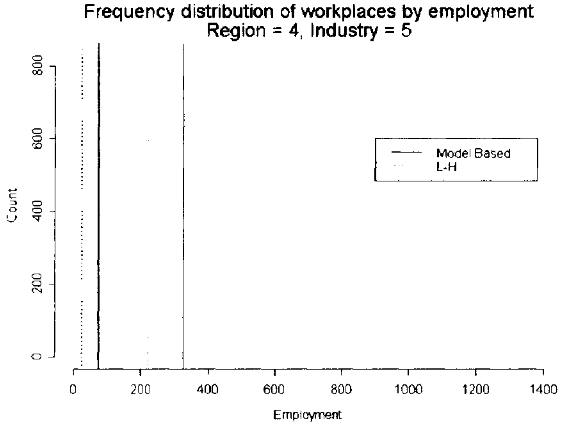
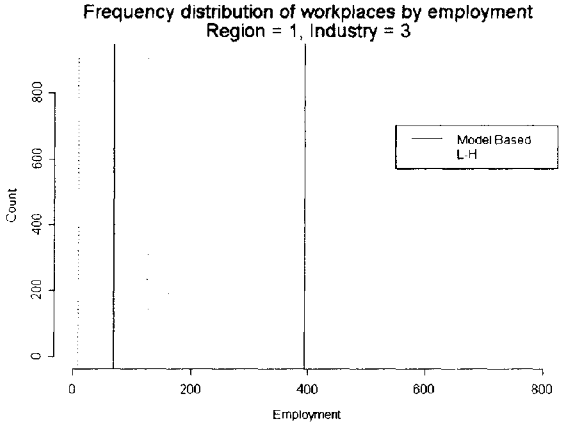


Figure 4. Stratum boundaries for model-based and Lavallée-Hidiroglou stratification.



In contrast, the MMB_2 and $AMMB_2$ only require that 3% or 10% of the sample be certainty units. This gives the survey methodologist the option of rotating out most of the sample while retaining only the very large units. The two graphs below show the difference in stratum boundaries between the MMB_2 and the LH techniques. The rightmost dotted line represents the certainty unit boundary determined by the LH method. The rightmost full line is the boundary for the stratum with the largest population units determined by the MMB_2 technique. Note that these units may be, and often are, selected with a probability less than one.

In both examples the LH moves the boundary for certainty units too far left to retain its optimal status. This results in a sample with a large proportion of non-probability units. Relaxing the optimality criterion leads to an MMB_2 sample, which still retains a high degree of efficiency while keeping the number of certainty units to a manageable size. It is these reasons that led us to apply the MMB_2 approach to WES.

3.2. Employee Portion

The target population is defined as all persons drawing pay for services rendered or for paid absences and for whom the in-scope employer must complete a Revenue Canada T-4 Supplementary Form. There is no up-to-date data source currently available containing a list of all employees in Canada, linked to their workplaces. A list of employees is created for each sampled workplace based on information obtained from the employer.

Once a selected workplace is contacted, a list of employees is created. This list contains all employees at the workplace including all paid on-site or off-site employees. Depending on the size of the workplace a systematic sample of three, six, nine, or twelve employees is drawn from the list by the interviewer using instructions provided by methodology. A benefit of this differential sampling of employees is that it results in potentially more representative occupational group coverage.

4. COLLECTION, EDIT, OUTLIER DETECTION, AND IMPUTATION

Data for both portions of WES are collected in seven Statistics Canada regional offices. CAPI and CATI applications have been designed for employer and employee data.

4.1. Data Collection and Edit

Prior to dispatching an interviewer to the workplace a telephone contact is initiated. It serves to identify a primary respondent within the workplace, typically the human resource person, and to collect basic tombstone information about the business. During this stage each unit is classified as being in-scope (i.e., has paid employees) or not out-of-scope (i.e., has no paid employees). Out-of-scope units are flagged and not interviewed.

The workplace questionnaire contains ten distinct blocks. Each block focuses on a different theme. In most cases a single respondent will be able to answer all the questions. If the primary respondent is unable to provide the requested information in its entirety, then he or she will be asked to identify the person privy to this information. The capture vehicle is capable of accepting up to ten different respondents, one for each content block. In addition to basic contact descriptors, the system will also store data on the respondent's position within the workplace. These data will be analysed to shed some light on how information is propagated within the management structure of the workplace.

The CAPI capture vehicle performs validity, range, and interfield edits. These are the types of edits that are performed during the collection of the first wave data. For subsequent waves we will develop a suitable suite of historical edits.

The majority of inter-field edits are confined to a single content block. If an edit failure occurs between blocks, then the primary respondent is asked to confirm the information.

An example of a validity edit is that total annual expenditures are positive. The corresponding range edit requires that expenditures do not exceed a large upper bound. A related interfield edit for total annual expenditures ensures that the sum of all expenditure sub-components such as wages and salaries, training expenses, and cost of implementing new technologies does not exceed total annual expenditures.

Upon the completion of the interview, the primary respondent will be asked to provide a list of employees attached to the workplace. Depending on the size of the employer, three, six, nine, or twelve employees will be selected systematically from the list. Given some basic personal employee data, the interviewer will print out a personalised employee participation form containing six mandatory questions. The primary respondent is then asked to distribute these forms amongst the selected employees.

Only the participation form is mandatory. The main employee questionnaire is voluntary. If the employee agrees to take part in the survey, then the corresponding regional office will contact him or her for a telephone interview. Employees failing to return the participation form will be followed up through their employer. This places response burden on the primary respondent in the workplace, and is the reason for selecting at most twelve employees from the workplace. The employee CATI application performs validity, range, and interfield edits. Any edit failures are resolved during the telephone interview.

4.2. Outlier Detection and Treatment

The use of CATI and CAPI for data collection greatly reduces the number of gross response and typographical errors. If either type of error remains undetected, then a *multivariate outlier detection* routine, the modified Stahel-Donoho approach, is applied to all complete and partial respondents prior to imputation. The method uses robust Mahalanobis distance, $MD_i^2 = (\mathbf{x}_1 - \overline{\mathbf{x}}_{rob})^T \Sigma_{rob}^{-1} (\mathbf{x}_1 - \overline{\mathbf{x}}_{rob})$, to identify units for which this statistic exceeds a prespecified percentile of the corresponding χ^2 distribution. The variable $\overline{\mathbf{x}}_{rob}$ in MD_i^2 is the L_1-estimator of multivariate location (Rousseeuw and Leroy, 1987), Σ_{rob} a robust estimator of scatter computed using a method similar to projection pursuit, and $\mathbf{x}_i$ is a vector of responses for unit i. This type of outlier detection is performed for both workplaces and employees at the micro data level. The sensitivity of the process can be adjusted to suit the survey's needs.

The current implementation of the outlier detection routine does not incorporate design weights. To be able to use the technique successfully with business survey data, one has to satisfy two criteria: (a) data homogeneity, and (b) data symmetry. Achieving data homogeneity obviates the need to use design weights when pooling neighbouring strata to increase the resolution of the outlier routine. Data homogeneity reduces the effect of the design and the complex problem of identifying aberrant observations in a sample drawn from a finite population reduces to a much simpler problem of dealing with outliers in the context of an infinite population.

Homogeneity can be achieved by applying an appropriate function to one or more variables. After data have been suitably transformed (eg., square root, log, etc.), the distributions of the resulting variables should be evaluated for approximate symmetry. This requirement stems from the fact that most outlier detection theory has been developed for contaminated normal distributions. The modified Stahel-Donoho approach is no exception. For WES approximate symmetry is achieved for ratios of continuous variables to total employment.

The outlier routine can be applied to respondents of a single wave, or across waves. To do so, the response vector $\mathbf{x}_i$ would be modified to include data from two consecutive waves. The possibility of extending the utility of the approach beyond two waves will be studied shortly. Our goal is to develop a method that would fill the gap between cross-sectional outlier detection and robust time series analysis.

Data validation is also performed at a macro level. For a number of key variables we identify the top ten contributors to the weighted estimates for further analysis.

Subject matter officers identify both micro and macro level anomalies and correct errors. After errors have been corrected, the data validation cycle is repeated. All remaining outliers are flagged and excluded from imputation.

4.3. Imputation

Imputation methods are used cross-sectionally for item non-response for units appearing within each wave for the first time. Longitudinal imputation methods are for wave non-response if historical data are available. In the absence of prior information, total non-response is handled by modifying the weights of the respondents. This approach assumes that the non-response is occurring completely at random.

There are three imputation methods being used for the first wave of the employer portion of WES: deterministic, distributional, and weighted hot deck. Deterministic imputation is used when a single missing field can be deducted from the given information. For example, if one component of a sum is missing and the remaining components including the sum are present, then the missing component can be determined uniquely.

Distributional imputation is used for questions where the respondent is asked to provide a total and its breakdown into multiple categories when either two or more of the categories are missing. The distribution of the categories is computed at a macro level and applied at the micro level. To illustrate this approach, let us assume that the respondent gave us total employment but was unable to provide a breakdown by occupational group. We would apply the distribution of the occupational groups computed at the industry/size level to the total employment figure to impute the missing fields.

For weighted hot deck, a missing field is imputed using the response of a suitable donor. The donor is selected randomly. The probability of selection is equal to the ratio of its sample weight over the sum of the sample weights of units in the corresponding cross-sectional imputation class. The weighted hot deck approach was adopted for the following four reasons. The method is easy to implement. It leads to approximately p unbiased point estimates (Rao, 1996). A consistent variance estimator can be constructed in the presence of imputed data (Rao, 1996). And lastly, most questions are independent keeping the number of post-imputation adjustments to maintain internal data consistency to a minimum.

Missing data on the employee questionnaire are imputed using nearest-neighbour imputation. Other imputation methods were studied but all item-by-item approaches led to internal inconsistencies. For example, one could impute five hours for a person using a computer in an average day and five hours for using another technological device even though the individual works only eight hours a day. Since many such dependencies exist in the employee data, a post-imputation system would have to have been very sophisticated to maintain all the inter-field relationships.

5. WEIGHTING AND ESTIMATION

Estimation for the Workplace and Employee Survey is also divided into two parts, a workplace portion and an employee portion.

5.1. Workplace Portion

The workplace portion of WES is a stratified single-stage simple random sample of workplaces drawn without replacement. Population estimates are computed using the separate ratio estimator. This improves the efficiency of the estimates by adjusting sample totals to known population totals available from the Business Register. The auxiliary information, total employment, is applied at the stratum level. Sparse strata are collapsed to improve the stability of the calibration process.

We would like to introduce some notation before formulating the estimation process. Note that the superscript t denoting time is implicit in all the following expressions. Let U_{1h} be the population of workplaces of size N_{1h} in stratum h, where $h = 1,\dots,L$; let s_{1h} be the corresponding intersection of the population with the sample of size n_{1h}. Stratum is defined as the crossing of industry, region, and employment size.

Let y be the variable of interest and let y_{1hi} be its value for the i^{th} unit in stratum h in the first stage of sampling (workplace stage). Then the estimated total for a specified domain of interest U_d is given by

$$\hat{Y}_{1,RAT}(d) = \sum_{h=1}^{L} \hat{Y}_{1h}(d) \frac{X_{1h}}{\hat{X}_{1h}},$$

where $\hat{Y}_{1h}(d) = \frac{N_{1h}}{n_{1h}} \sum_{i=1}^{n_{1h}} y_{1hi}(d)$, $X_{1h} = \sum_{i \in U_{1h}} x_{1hi}$, and $\hat{X}_{1h} = \frac{N_{1h}}{n_{1h}} \sum_{i \in s_{1h}} x_{1hi}$. The corresponding estimated variance is

$$\hat{V}(\hat{Y}_{1,RAT}(d)) = \left\{ \sum_{h=1}^{L} N_{1h}^2 \left(\frac{1}{n_{1h}} - \frac{1}{N_{1h}} \right) \frac{1}{n_{1h} - 1} \right\} \times \left(\frac{X_{1h}}{\hat{X}_{1h}} \right)^2 \sum_{i=1}^{n_{1h}} (e_{1hi}(d) - \overline{e_{1h}(d)})^2,$$

where

$$e_{1hi}(d) = y_{1hi}(d) - \tfrac{\hat{Y}_{1h}}{\hat{X}_{1h}} x_{1hi} \text{ and } \overline{e_{1h}(d)} = \tfrac{1}{n_{1h}} \sum_{i=1}^{n_{1h}} e_{1hi}(d).$$

In odd years, cross-sectional estimates will be unbiased and representative of the corresponding cross-sectional population. In even years, by virtue of not sampling birth units, the cross-sectional estimates will be somewhat biased. Post-stratification and other techniques will be used to keep the bias to a minimum.

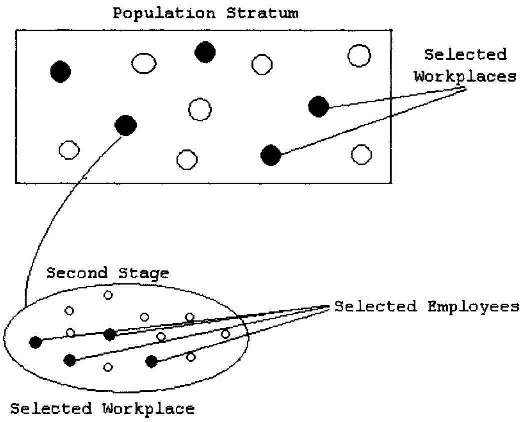
5.2. Employee Portion

The employee portion of WES is a stratified 2-stage design with workplaces drawn at the first stage and employees drawn at the second stage. Within a given workplace (defined by the index hi), we sample systematically n_{2hi} out of N_{2hi} units. This results in a sample of employees: $s_2 = \bigcup_{h=1}^{L} \bigcup_{i=1}^{n_h} s_{2hi}$. The measurement on an individual is denoted by y_{2hij}, where the leading term of the index is used to specify the stage of sampling.

The estimated total for a given domain U_d has the form

$$\hat{Y}_2(d) = \sum_{h=1}^{L} \frac{N_{1h}}{n_{1h}} \left(\frac{X_{1h}}{\hat{X}_{1h}} \right) \sum_{i=1}^{n_{1h}} \frac{N_{2hi}}{n_{2hi}} \sum_{j=1}^{n_{2hi}} y_{2hij}(d) .$$

Figure 5. Sampling of employees.



The overall estimated variance of $\hat{Y}_2(d)$ can be written as the sum of the variances for the two corresponding stages, I and II,

$$\hat{V}(\hat{Y}_2(d)) = \hat{V}_I(\hat{Y}_2(d)) + \hat{V}_{II}(\hat{Y}_2(d)) . \qquad (5.1.)$$

The first component of (5.1.) is defined as

$$\hat{V}_I(\hat{Y}_2(d)) = \left\{ \sum_{h=1}^{L} N_{1h}^2 \left(\frac{1}{n_{1h}} - \frac{1}{N_{1h}} \right) \frac{1}{n_{1h}-1} \right\} \times \left(\frac{X_{1h}}{\hat{X}_{1h}} \right)^2 \sum_{i=1}^{n_{1h}} (e_{1hi}(d) - \overline{e_{1hi}(d)})^2 ,$$

where $e_{1hi}(d) = \hat{Y}_{2hi}(d) - \overline{\hat{Y}}_{2h}(d)$. The two terms in the residual expression are $\hat{Y}_{2hi}(d) = \frac{N_{2hi}}{n_{2hi}} \sum_{j=1}^{n_{2hi}} y_{2hij}(d)$ and $\overline{\hat{Y}}_{2h}(d) = \frac{1}{n_{1h}} \sum_{i=1}^{n_{1h}} \hat{Y}_{2hi}(d)$.

The second component of (5.1.) has the form

$$\hat{V}_{II}(\hat{Y}_2(d)) = \left\{ \sum_{h=1}^{L} \frac{N_{1h}}{n_{1h}} \sum_{i=1}^{n_{1h}} \left(\frac{X_{1h}}{\hat{X}_{1h}} \right)^2 N_{2hi}^2 \left(\frac{1}{n_{2hi}} - \frac{1}{N_{2hi}} \right) \right\} \times \frac{1}{n_{2hi}-1} \left\{ \sum_{j=1}^{n_{2hi}} (y_{2hij}(d) - \bar{y}_{2hi}(d))^2 \right\}$$

where $\bar{y}_{2hi}(d) = \frac{1}{n_{2hi}} \sum_{j=1}^{n_{2hi}} y_{2hij}(d)$.

The preceding formulation of variance is that for simple random sampling without replacement. If the ordering of individuals drawn from the employee lists provided by the selected employers is assumed to be random, then SRSWOR can be used as an approximation for systematic sampling.

6. FUTURE DEVELOPMENT

The preparations for the first wave are well under way. Stratification, sample allocation, sample selection, collection vehicle, edit and imputation, outlier detection, weighting and estimation have been put in place. A small live test will be conducted in August of 1998 to test the new systems that have been developed since the large-scale pilot. This small test sample will also be used in the future to test different longitudinal strategies.

A small study will be started shortly to identify the best method for longitudinal imputation. It is conceivable that different imputation strategies may be developed for small, medium and large employers. Currently there are no plans for backward revision of weights and estimates. Under the proposed scenario the weights of responding units will be modified to adjust for total chronic non-response. This is assuming that the total non-response is occurring completely at random (Little and Rubin, 1987) and that we will not be able to convert refusals.

Even prior to conducting an in-depth analysis of the non-response patterns, there are some options available to us that we will explore. The possibility exists that longitudinal weights, along with the corresponding estimates, may be revised after a respondent has decided to stop providing survey data. Revisions may also be necessary if a refusal is converted and used to impute his or her past. We will also examine data collected by other surveys to determine whether cross-imputation can be used to impute some missing items in WES. These are some of the questions that will be resolved in the coming months.

Stratification is currently using estimated employment available on the Business Register to stratify the WES population by size. The employment estimates may not be updated on a regular basis, thus making the size variable stale. Recently an agreement has been reached with Revenue Canada whereby, in addition to monthly remittances, we will receive current employment figures for all live employers. This will not only improve stratification by size but it will also reduce the number of potential stratum jumpers. This will in turn increase the efficiency of survey estimates.

The number of potential stratum jumpers is reduced by initially stratifying the population using true employment figures rather than the corresponding estimated quantities available from the Business Register. Since he majority of changes result from using a proxy, rather than the real variable of interest, for stratification, this approach will virtually eliminate the occurrence of fictitious stratum jumping.

7. ACKNOWLEDGEMENTS

The authors wish to thank D. Yeo, G. Laflamme and B. Lapierre for their helpful comments and suggestions during the writing of this paper.

8. REFERENCES

Godfrey, J., A. Roshwalb, A. and Wright, R. L (1984). Model-Based Stratification in Inventory Cost Estimation. *Journal of Business & Economic Statistics*, Vol. 2, No. **1**, 1-9.

Lavallée, P. and Hidiroglou, M. A. (1988). On the Stratification of Skewed Populations. *Survey Methodology Journal*, Vol. 14, 33-44.

Little, R. J. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Mahalanobis, P. C. (1952). Some aspects of the design of sample surveys. *Sankhya* **12**, 1-7.

Rao, J. N. K. (1996). On Variance Estimation With Imputed Survey Data. *Journal of the American Statistical Association*, Vol. 91, No. 434, pp. 499-520.

Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley & Sons.