ON REDESIGNING CANADA'S ESTABLISHMENT BASED EMPLOYMENT SURVEY

David Dolson, Statistics Canada
11-J R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6

KEY WORDS: Administrative data, regression estimation, two-phase sampling

1. The Current Survey

Canada's monthly establishment survey to measure the volume of employment was established in 1918. Its most recent redesign, implemented in 1983, is called the Survey of Employment, Payroll, and Hours (SEPH). It collects data on payroll employment, weekly earnings, and weekly paid hours. The primary objectives currently include:

> to provide monthly estimates of the total number of paid employees, average weekly earnings, average weekly hours and other related variables at the industry division by province level.

> to provide these estimates for Canada at the three digit Standard Industrial Classification (SIC) level

The list of establishments SEPH uses as its frame is derived from Statistics Canada's business register (BR). For each monthly survey cycle the frame is updated for births, deaths etc. as reflected on the BR. The primary source of information for maintenance of the BR is the Payroll Deduction (PD) accounts each employer has with Revenue Canada. A group of establishments linked together by ownership or control is called an enterprise. On the BR, each PD account is linked to the enterprise to which it belongs. It is primarily through the births, deaths etc. of these PD accounts that the BR is maintained. A more detailed discussion of the BR is given by Cuthill (1989).

SEPH covers all industries except agriculture, fishing and trapping, private household services, religious organizations, and military services. It is designed as a stratified sample of establishments with stratification by industry division (16), province or territory (12), and employment size group (4). Each stratum is further subdivided into sub-strata by 3 digit SIC called cells. The sampling within each cell is simple random without replacement.

The required precision of the estimate of total employment is specified at the industry division by province level. To achieve this, a sample of about 60,000 establishments is selected from the population of about 800,000. Of these, about 27,000 are self-representing; these are primarily establishments belonging to enterprises having 200 employees or more. The remaining take-some sample is allocated to strata in proportion to the estimated number of employees in the take-some population of each stratum. Within each stratum the sample is further allocated to cells in proportion to the population size in each cell. The take-some sample is rotated at the cell level. Sampled units remain in the sample for at least a year, except for sampled births which generally remain in the sample for fewer occasions. Units which rotate out of the sample are kept out for at least a year.

Dead units detected by independent sources as well as from SEPH are removed from both the sample and the frame in order to simplify their treatment operationally. To maintain a nearly unbiased estimator, an estimate of dead units in the population, called the death adjustment factor, is used when computing weights. Schiopu-Kratina and Srinath (1991) have shown that the resulting estimator performs better, conditionally, than other more traditional estimators of totals.

Estimation of totals and variances is done at the cell level and these estimates are aggregated to the desired level. The basic structure for the SEPH estimator of total is $\hat{N}\overline{X}$ where $\hat{N}$ is the estimated number of live units in the population and $\overline{X}$ is the mean of the variable. $\hat{N}$ is not allowed to exceed the actual number of units in the population which may include some unknown deaths. Although this estimator is not unbiased it has a smaller mean squared error than the unbiased estimator in which $\hat{N}$ is unconstrained.

A more extensive description of the current SEPH methodology is given by Schiopu-Kratina and Srinath (1991).

2. Redesign Considerations

A secondary objective of the current SEPH is to produce estimates at the three digit SIC-province level annually. This as well as the second primary objective led to the choice of a detailed stratification using 214 three digit SIC industries in the current design. In fact, SEPH produces three digit SIC-province estimates

monthly. However, detailed estimates such as these are now more clearly viewed as having much less importance than larger aggregates.

The primary objectives of the redesigned survey will include production of good quality estimates of total payroll and total employment each month at the national level by three digit SIC and provincial level by industry division. At a more detailed level emphasis will be placed only on a few "important" three digit SICs in each province. The estimates of total payroll are especially important for estimation of monthly labour income.

Hence there will be a reduced need for detailed industrial strata. "Important" industries may be identified as design strata, while other industries will not. Estimates for these latter industries will likely be less precise than those for the "important" industries.

A result of the detailed stratification in the current design is many cells with small populations (and small employment) but large sampling fractions. The sample rotation methodology developed to cope with this is complex, as is the computer system which implements it; simpler methods are now available and a less detailed stratification will alleviate the problem.

There was a cost to use of the death adjustment factor (daf). Neither the methodology nor its computer systems implementation are simple and it has been problematic in survey maintenance. So, at the cost of a small loss of efficiency, it is now desired to simplify the methodology and systems related to the treatment of deaths by removal of the daf and moving to a more standard treatment of deaths.

Like many surveys of its generation SEPH overedits its data. A large part of the survey budget is spent in this operation. SEPH's editing will be improved and its cost reduced by using newer methodology and systems. In particular, selective editing methods proposed by Hidiroglou and Berthelot (1986) and by Latouche and Berthelot (1992) will be used. These are being done using Statistics Canada's generalized DC2 system for data collection, capture and edit.

Generalized survey processing software was not readily available when SEPH was being designed and SEPH specific systems were developed. Some of these systems have turned out to be excessively rigid and hard to maintain. There are several new developments in computing hardware and software that are being used to develop generalized systems at Statistics Canada which will be helpful in a redesigned SEPH. Computer Assisted Interviewing methods can reduce costs while improving data quality and timeliness; this methodology will be used for some of the data collection in the redesigned survey. The generalized software for survey processing being developed at Statistics Canada that will be used for SEPH are: DC2 as noted above; GEIS, the Generalized Edit and Imputation System; and GES, the Generalized Estimation System.

Finally, and most important, some new administrative data are now available from the payroll deduction data source. Those employers who are to remit payroll deductions monthly to Revenue Canada are now asked to report on the PD7 form which accompanies their payment, the *total payroll for the month* and the *total number of employees* for the last pay period of the month. These monthly remitters are generally smaller employers. Larger employers, who make remittances more often, are not currently required to provide these data on their remittance forms.

Because these new data are available for a large fraction of SEPH's target population a substantial reduction in the SEPH sample size, and hence in the cost of the survey, will be possible. This factor in particular, has provided the impetus to redesign the survey. All necessary redesign activities are scheduled for completion so that the reduced sample size can be implemented for the survey with January 1994 reference month.

3. The Redesigned Survey

Because the new administrative data are available only for the smaller businesses which are required to supply the two new variables on their PD7 forms, SEPH will use two frames - the ESTABLISHMENT frame consisting of a list of establishments, and the ADMIN frame consisting of a list of PD accounts. They are derived from the Business Register and the list of all PD accounts.

Any enterprise which has at least one PD account for which the new data are not required has all of its establishments placed in the ESTABLISHMENT frame. In addition, all establishments belonging to enterprises having more than one establishment or more than 99 employees are also included. All PD accounts for such enterprises, whether the new data are required or not, are therefore excluded from the ADMIN frame. The ESTABLISHMENT frame will include about 100,000 establishments accounting for

about 70% of total employment. A monthly survey similar to the current SEPH will be designed for this frame.

The ADMIN frame includes all remaining PD accounts required to supply the new variables. It will comprise about 800,000 PD accounts, accounting for about 30% of total employment. In the short term (two to three years) it is too costly to capture the new data for all accounts every month. (In the longer term Revenue Canada will automatically capture the data for all of these accounts and provide them to Statistics Canada). Consequently a two-phase sample will be selected. The first phase sample of PD accounts, for which data will be captured from the PD accounts, is called the ADMIN sample. From this sample, a subsample will be selected to collect data for the other SEPH variables not available on the PD7 forms.

3.1 The ESTABLISHMENT Survey

The ESTABLISHMENT frame will be stratified by province (12), industry set and employment size. The industrial stratification will be province specific and oriented towards "important" industries within the province -- generally those with large employment. SEPH subject matter experts initially identified 740 such industry province combinations. Each of these may constitute an industry set for the given province; those with much of their employment or much of their variance coming from the non-self-representing part of the ESTABLISHMENT frame will be retained as industry sets. Remaining three digit SICs will be aggregated to one or more industry sets defined at higher levels of industrial classification so as to balance the need for adequate homogeneity in these strata with the need to constrain the number of them to a reasonable level. At the time of writing, the number of industry sets per province ranges from a low of 13 to a high of 54 for a total of 360. (This compares to 1863 in the current design.)

There will continue to be four levels of size stratification, uniform for all provinces and industries. All establishments, regardless of size, belonging to enterprises having 300 employees or more will be self-representing. This boundary of 300 employees is a compromise between a number of factors. First, for the purpose of allocation of estimates of labour income to industry and province, data are required from SEPH for complex structured enterprises operating in more than one three digit SIC or province. This boundary will include with certainty enterprises accounting for the large majority of earnings of complex structured enterprises. Secondly, it is also a compromise between the needs of generally smaller industries and provinces where a lower take-all boundary would be more optimal and those of bigger industries and provinces where a higher boundary would be better. The total sample of 31,000 for this frame will consist of about 21,300 self-representing establishments plus about 9,700 establishments selected from the non-self-representing population of about 78,700.

A more efficient approach would implement a design with industry-province specific employment size stratification, including the take-all boundary. However, to meet the January 1994 implementation date we are constrained to simple modifications of our existing system which requires that the same employment size stratification be used in all provinces and industries. A province specific approach may be implemented at a later date.

Sample allocation will be determined via an approach which initially specifies a target coefficient of variation for estimated total employment for Canada. Then, this will be translated to a CV target for estimated total employment for each province; these targets will vary to a limited extent between provinces. Within each province a CV target is then derived for each industry set. Finally this translates into a CV target for the ESTABLISHMENT portion by adjusting for the CV of estimated total employment for the ADMIN portion. This approach is described in a more generalized context by Latouche (1988).

Sample selection and rotation will continue as it does currently with one exception. It will be simplified by removal of the death adjustment factor. Instead, dead units detected by the survey will be retained in the sample until they would normally rotate out. In the longer term SEPH plans to move to a newer and simpler sample rotation method like the modified collocated sampling strategy described by Srinath and Carpenter (1993).

In general, estimation of totals will continue to use the expansion type estimator currently used by SEPH. However, when estimates are needed for industries not separately identified as strata, post-stratification will be used. The use of a sample size dependent estimator is also being considered for small domains. Several of these are described by Srinath and Hidiroglou (1985).

3.2 The ADMIN Survey

The ADMIN sample (10% of the frame in the three largest provinces, 100% in the two territories, 20% elsewhere) is manually selected each month and is a systematic sample of PD account numbers. This sample has been in place since January 1993. Although deaths are deleted and births added, no sample rotation takes place. From this sample, data are available for total employment and total payroll (these being the two new variables added to the PD accounts) but not for the full range of SEPH variables. Starting in January 1994, a subsample of 7,500 will be selected from those accounts on the frame which are potentially alive and classified for both industry and province to collect data for these other variables.

3.2.1 The ADMIN Sample

A first step in the processing of the ADMIN sample is its treatment for missing data. In any given month it is expected that no PD7 form will be received for about 30% of accounts. A large fraction of this is accounts for which there are no employees in the month due to temporary or seasonal closure; such employers are requested not to send in their PD7 forms. For very many of these units, it is known apriori that no remittance is expected and codes are maintained indicating this; imputation of zero employment and payroll for such units is easy. Employers who do have employees but for whom the PD7 form is not received in time and those who send in their PD7 forms but fail to indicate either or both of total employees and total payroll will be considered as non-respondents. Finally many deaths may (initially) be indistinguishable from non-response by a live unit. For these latter two groups, deterministic imputation is done when information is available for the same units from the previous month and using averages and trends for imputation groups (generally two digit SIC by province group combination). When such information is not available, a weighting adjustment is made.

3.2.2 The ADMIN Subsample

From the ADMIN sample, data will be available for total employment and total payroll but not for the full range of SEPH variables. To collect data for these other variables a subsample of 2,500 will be contacted each month. They will be selected from those accounts on the frame potentially alive and classified for both industry and province. However, this very small monthly sample which our budget and response burden considerations allows us is not considered to be adequate and it is planned to "borrow strength" temporally to improve the estimates. Although more sophisticated methods are available, it is planned to adopt a relatively simple one, as follows.

A subsample of 7,500 PD accounts will be selected. Rotation as well as updates for births and deaths will take place every month with each sampled unit being kept in the sample for at least one year followed by at least one year out of sample. It will be split up into three portions of 2,500, each representative by industry and province. One portion will be surveyed each month and each portion will be resurveyed quarterly. At the estimation stage each month, data for the full sample of 7,500 will be used by combining the sample for three consecutive months, centred at the month in question.

Like the ESTABLISHMENT sample, the ADMIN subsample will be stratified by province, industry and employment size group. Again, the industrial stratification will be oriented towards "important" industries. Because of the very small sample size, the stratification may have to be at a more aggregated level than that for the ESTABLISHMENT frame.

Where possible, the ADMIN subsample will be stratified by employment size group. This stratification will likely have at most two levels - 0-19 employees and 20 or more. Only one level will be used in situations where the population or expected sample size is too small. The small units covered by the ADMIN frame have very dynamic employment levels. Thus more levels of employment size stratification will likely be avoided in order to minimize difficulties with stratum jumpers.

The purpose of the ADMIN subsample is for estimation of total hours and the allocation of hours, earnings and employment to categories of employee (paid by the hour, salaried, other). Sample allocation will be oriented towards maximizing the efficiency of estimates of total hours.

3.2.3 Estimation for the ADMIN Frame

For total employment and total payroll for the month, estimation can proceed directly, using the sample of PD accounts. For all other variables, a model assisted approach will use information from both the sample and the subsample. For an excellent discussion of model assisted methods, see Särndal, Swensson and Wretman (1992).

Model groups consisting of sets of strata from the subsample will be defined. Normally a model group will consist of a number of industry sets within a province. In some cases where subsample sizes will are too small a model group may cover more than one province for its industry set(s). Regression estimation will be done at the level of the model group using total employees and total payroll for the month as the independent variables. For each model group, estimates for these two variables are controlled to be equal to the direct estimates from the ADMIN sample.

In the near term, a specific model assisted method described in section 7.12 of Särndal, Swensson and Wretman (1992) will be used; observed values are used for units in the subsample and predicted values for remaining units. This estimator, called a cosmetic estimator, is also discussed by Särndal and Wright (1984). It will be implemented via mass regression estimation. Within the context of a broader discussion of imputation, this procedure is discussed by Kovar and Whitridge (1993). Regression parameters will be estimated for each model group using the subsample data. Values for all of the other SEPH variables will be imputed for each PD account in the ADMIN sample but not in the subsample, model group by model group, using the appropriate estimated regression parameters. Although this procedure is unbiased for model groups, it is potentially biased for domains below the model group level if the model fails. This procedure also has the property that estimates of the other SEPH variables for small domains which are not represented in the subsample will be synthetic. In order to minimize the risk or frequency of negative imputed values that may occasionally arise, model groups will have to be sufficiently large as to ensure an adequate sample size while not so large as to be non-homogeneous with respect to the assisting model. Variance estimates will be available for total employees and for total payroll for the month, but not for the other SEPH variables due the use of mass imputation.

In the longer term, it is hoped to implement estimation via a modified version of the generalized regression estimator using the Generalized Estimation System (GES) being developed at Statistics Canada. Model groups will define the level at which the regression is carried out. The ADMIN data from the sample will be linked to specific model sub-groups and computation of g-weights will account for these data at this level.

The frames for a given reference month m, are first constructed in m-1 and are based upon information as of the end of m-4. The ADMIN sample, which is selected and captured in m+1, will include not only accounts on the frame but also new accounts from m-3, m-2, m-1, and m. The frames will be updated to include these units whether sampled or not. Those new accounts belonging to enterprises covered by the ESTABLISHMENT frame will be dropped while the remaining ones will be added to the ADMIN frame.

For reference month m, preliminary estimates are published in m+2 with revised estimates in m+3. At this time the ADMIN frame consists of three sets of accounts -- those which were eligible for selection to the subsample, newly classified units (both new and old), and unclassified units (both new and old).

For the unclassified units, all that can be done is to estimate their total employment and total payroll. The other SEPH variables cannot be estimated since these units are not represented in the subsample in any way.

The newly classified units were not eligible for inclusion in the subsample. However, they will be included in their appropriate industrial strata for estimation purposes as if they had been eligible. This is not a problem for estimation of total employment and total payroll where the data come from the sample. However, for estimation of the other SEPH variables it assumes that the relations between variables are not different from those for units which were eligible for selection into the subsample. SEPH subject matter experts believe this to be a reasonable assumption. Further, it is believed that even if not true, the bias will be small and acceptable since it would affect only a small fraction of the population and only in the distribution of estimated total employment, total payroll and total hours to various categories.

Estimation for reference month m will be carried out using data collected for reference months m-1, m, and m+1. From a collection point of view, although data collection will be slower for the larger units from the ESTABLISHMENT frame - for whom collection is primarily by mail - it is expected that the CATI collection for the ADMIN subsample will provide m+1 data early enough to be usable for estimation for reference month m. From the estimation point of view, this procedure assumes temporal stability - over three months - of the assisting model. In a few highly seasonal industries this is believed to be a poor assumption. In these cases there is a trade-off between variance on one hand and model bias on the other.

Bias can be reduced by a procedure in which reduced "weight" is given to the data from m-1 and m+1 at the cost of increased variance due to a reduced effective sample size. It is important to note that this does not affect estimation of the primary variables, total employment and total payroll, and is applied only in the ADMIN frame, affecting on average estimates covering about 30% of employment.

A final stage of estimation is combining estimated totals from the ESTABLISHMENT portion and the ADMIN portion to produce estimated totals for the entire target population. At this point ratios such as average weekly earnings, average hourly earnings etc. can be computed.

4. Concluding Remarks

The new data available from the administrative source allows for a significant improvement in the estimates for total employment and for total payroll while reducing the respondent burden amongst small businesses. Because the frame can be updated to include the most recent births, SEPH estimates will reflect a more current population than the current survey. The survey design will be more efficiently oriented to industries which are most important. Although a thorough discussion is out of scope for this paper, SEPH is making major improvements to its data collection, capture, edit and imputation procedures which will reduce the survey's costs and help improve its data quality. In the medium term SEPH will also be simplifying its sample selection and rotation procedures, possibly using Statistics Canada's Generalized Sampling System, GSAM that is under development. As well, some aspects of estimation are being implemented using the Generalized Estimation System, GES. The new survey will be less costly, more efficient, more flexible, easier to maintain and produce improved data quality.

This paper describes SEPH redesign plans as of July 1993, but since the redesign is still under way these plans remain subject to change.

ACKNOWLEDGEMENTS

I would like to thank B. Armstrong, B.N. Chinnappa, D. Binder, M. Latouche, P. Lys, and K.P. Srinath for their helpful comments.

REFERENCES

Hidiroglou, M.A. and Berthelot, J.-M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys," *Survey Methodology*, 12, 73-83.

Cuthill, I.M. (1989), "The Statistics Canada Business Register," *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 69-86.

Kovar, J.G. and Whitridge, P.J. (1994), "Imputation of Establishment Survey Data," in B.G. Cox et al (eds.) *Survey Methods for Businesses, Farms and Institutions*, New York: Wiley.

Latouche, M. (1988), "Sample size determination, allocation and selection," *Methodology Branch Working Paper BSMD-88-021E/F*, Ottawa: Statistics Canada.

Latouche, M. and Berthelot, J.-M. (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys," *Journal of Official Statistics*, 8, 389-400.

Särndal, C., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Särndal, C. and Wright, R. (1984), "Cosmetic form of estimators in survey sampling," *Scandinavian Journal of Statistics*, 11, 146-156.

Schiopu-Kratina, I. and Srinath, K.P. (1991), "Sample rotation and estimation in the Survey of Employment, Payroll and Hours," *Survey Methodology*, 17, 79-90.

Srinath, K.P. and Carpenter, R. (1994), "Sampling methods for repeated business surveys," in B.G. Cox et al (eds.) *Survey Methods for Businesses, Farms and Institutions*, New York: Wiley.

Srinath, K.P. and Hidiroglou, M.A. (1985), "Sample Size-Dependent Estimators for Small Areas With Applications to Business Data," in R. Platek and M.P. Singh (eds.) *Small Area Statistics: An International Symposium*, 118-134, Ottawa: Carleton University Press.