Methodological Issues in Canada's Workplace and Employee Survey

Michel Hidiroglou, Pierre Lavallée, Zdenek Patak and Don Royce
Statistics Canada, Business Survey Methods Division
11th Floor, R.H. Coats Building, Tunney's Pasture
Ottawa, Ontario, K1A 0T6, Canada
Contact: don.royce@statcan.ca

1. Introduction

The Workplace and Employee Survey (WES) was developed by Statistics Canada to provide an integrated view of the activities of employers and their employees. The survey covers technology adoption, innovation, human resource practices, labour turnover and business strategies of employers, as well as wages, training, technology use, working hours and other workplace activities of employees. In addition to providing cross-sectional linked employer-employee data, the survey is also longitudinal. This will allow researchers to study both employer and employee outcomes over time. The first wave of the survey took place in 1999.

2. Linking data on employers and employees

Traditional business surveys focus only on the enterprise or some component thereof, e.g., the establishment. For the purposes of WES, both the work environment and the employee, and how they interact, are of interest. After considerable testing, the individual workplace was chosen as the unit of analysis for the employer portion of the sample, and the employee at the workplace was chosen as the unit of analysis for the employee portion.

2.1 Employer portion

More precisely, the target population for the employer portion was defined as all workplaces operating in Canada with paid employees, excluding workplaces in the three northern territories of Canada, and excluding workplaces in certain industries (agriculture and related industries, fishing and trapping, etc.) Businesses with no paid employees are out of scope, although this is sometimes not known until after the business is sampled and contacted.

The survey frame for workplaces was drawn from the Business Register of Statistics Canada, which maintains information on all Canadian businesses. The population of workplaces was stratified by industry, geographical region, and employment size. The method used for size stratification was a modification of a model-based approach due to Godfrey, Roshwalb and Wright (1984). Neyman allocation was used to allocate the sample to size strata within industry by region cells. All strata with sampling fractions of 80% or more were converted to take-all strata.

The sample sizes were computed to yield pre-specified coefficients of variation (9%) for the employment size variable at the industry by region level. A simple random sample of 11,719 workplaces was selected across some 252 strata. Of this number, some 2575 were retained as a contingency reserve. The remaining 9,144 workplaces were subjected to pre-contact, where dead and out-of-scope units (e.g., non-employer, excluded industry) were identified, and the primary respondent within the workplace was identified, typically the human resource officer. At the conclusion of pre-contact in April 1999, 7,932 potential respondents had been identified.

Following pre-contact, eligible respondents were mailed a questionnaire and then followed up by Computer Assisted Personal Interviewing (CAPI). The questionnaire contained ten blocks of questions, each focussing on a different theme. The CAPI vehicle conducted validity, range and some limited inter-field edits. Upon completion of the interview, the primary respondent was asked to provide a list of employees associated with the workplace, to be used to select the sample of employees.

Processing of the employer data consisted of the following three modules: multivariate outlier detection, imputation and estimation.

Multivariate outlier detection: The method used was a modified Stahel-Donoho method (Stahel (1981), Donoho (1982)). This method computes a robust Mahalanobis distance statistic and compares it to a pre-specified percentile of the chi-square distribution to identify those records requiring imputation. The method is applied to both complete and partial respondents.

Imputation: Three methods were used, in the following order: deterministic imputation, distributional imputation and weighted hot deck imputation. Deterministic imputation was used when the missing or incorrect value could be determined exactly from other variables on the questionnaire. Distributional imputation was used for questions where the respondent provided a total but no distribution of its components. In this case the average distribution for respondents was applied to the non-respondent. A weighted hot deck donor imputation method was used for the remaining non-response. Imputation classes were formed and donors were selected randomly with a probability of selection equal to the ratio of a donor's sample weight to the total of the sample weights of all units in the imputation class.

Estimation: A separate ratio estimator at the industry by region level was used to estimate the parameters (totals, ratios) of interest for various domains. The auxiliary variable used was total employment, which was available from the much larger Survey of Payroll, Employment and Hours.

2.2 Employee portion

The target population for the employee portion of the survey was defined as all persons drawing pay for services rendered or for paid absences. At the time of the interview with the primary respondent in the workplace, the interviewer obtained a list of all associated employees. Depending on the size of the workplace, a sample of three, six, nine or twelve employees was selected from the list of employees, with larger workplaces receiving larger samples of employees.

Once the sample of employees was selected, the interviewer printed out a personalized employee participation form containing six mandatory questions. The primary respondent was then asked to distribute these forms to the selected employees. Employees who did not return these mandatory forms were followed up through the employer.

The main employee questionnaire itself was voluntary. Employees who agreed to be interviewed were contacted from a Regional Office and interviewed using Computer Assisted Telephone Interviewing (CATI). Like the CAPI application for employers, the CATI application for employees performed various range, validity and inter-field edits. Outlier detection and imputation for the employee questionnaire were very similar to the methods used for the employer questionnaire. There were no edits between the employee and the employer questionnaires.

Estimation for the employee portion reflected the resulting stratified two-stage design with workplaces drawn at the first stage and employees drawn at the second stage. Again, auxiliary

information on total employment was used to improve the first stage weights. As no additional auxiliary information was available at the second stage, the second-stage weight was simply the inverse of the second-stage selection probabilities of the employees.

3. Generating longitudinal data

Starting with Year 2 of the survey, which was implemented in early 2000, the emphasis shifted to the longitudinal aspect of WES.

3.1 Workplace portion

For the workplace portion of the survey, there was no selection of a sample of births between Years 1 and 2. For Year 3, however, a sample of births since Year 1 was selected. This cycle will be repeated on every odd-numbered survey occasion. In odd-numbered years, unbiased estimates of the corresponding cross-sectional population can be produced. In even-numbered years, by virtue of not sampling births, cross-sectional estimates will be biased. It is anticipated that the use of post-stratification will reduce this bias.

Starting with Year 2, data collection for even-numbered years is carried out by CATI. The edits in the second and subsequent years include several historical edits for variables deemed to change over time. The multivariate outlier detection method described in Section 2.1 has also been modified to include a historical dimension, by including data from two consecutive years in the vector used to compute the robust Mahalanobis distance. Longitudinal imputation is used for non-response if historical data are available.

3.2 Employee portion

Sampling of employees at the workplace is a major expense for the survey, since it requires the time of an experienced and well-trained interviewer. For this reason, sampling of employees is done only every second year. Employees who are still at the workplace receive the same questionnaire they received in the previous year. Employees who have left the workplace receive a special exit questionnaire. This permits the production of longitudinal estimates of employees for the original population.

In the third year, the sample of employees will be redrawn independently of the sample drawn in the first year. Testing has shown that it is too difficult to obtain lists of exclusively new employees to the workplace, so it is not possible to sample "birth employees" alone. Again, this pattern will continue on alternate years. Some employees will be re-selected simply by chance in Year 3 of the survey. These employees will provide longitudinal estimates for the original population for at least four years or until they leave the workplace or the workplace closes down. Collection of the employee questionnaire will be by CATI in all years. The methods used for outlier detection, imputation and estimation are all very similar to those described above.

4. Future issues in the methodology of WES

As of this writing, collection has been completed for the first two years of WES, in 1999 and 2000 respectively. Plans for collection of the third year are being finalized, and preliminary plans for Year 4, as described above, have been developed. Beyond this, however, further plans will depend on our experience with the first three years. We will be considering the following issues:

Sample design: At present, the sample design for workplaces is based on a fixed panel of workplaces, with births to the universe being sampled every second year. However we recognize

that the panel is subject to attrition over time. To reduce the possibility of a stratum being completely eroded over time, the minimum sample size for all strata was set to ten. If, despite this precaution, sample attrition substantially depletes a particular stratum, we may have to add a second panel for that stratum. If we find that sample attrition due to response fatigue is a serious problem, then we may also consider introducing sample rotation of the smaller units, although there are currently no plans to do so. A second concern with the use of a fixed panel design is the deterioration in the efficiency of the stratification over time as the business universe changes. We plan to examine the impact of re-stratifying the population of workplaces and re-drawing the sample after the fourth year, while maximizing the overlap between the two samples of workplaces.

Outlier detection: Further research is also planned in the areas of longitudinal outlier detection. We will be examining the feasibility of including more than two consecutive years of data in the distance function. Our goal is to develop a method that would fill the gap between cross-sectional outlier detection and robust time series analysis. As well, the current implementation of the method does not permit the inclusion of design weights. Inclusion of weights in outlier detection is a more general topic for future research at Statistics Canada, not just for WES.

Imputation: In the area of longitudinal imputation, we have conducted simulation studies of a number of methods. We have implemented several different methods in Year 2, including ratio imputation and weighted hot-deck, and we will evaluate these methods in the near future

Estimation: Further research is needed to decide how best to handle cross-sectional estimates for even-numbered years, when there is no sampling of births. We are currently considering using regression modeling to reduce the bias. For longitudinal estimation, we need further research into how to treat various combinations of response patterns over the various years.

New user demands: Finally, Statistics Canada has begun developments on a new Labour Cost Index (LCI). One of the approaches suggested has been to add some additional questions to the WES employer questionnaire to collect some of the information necessary to produce such an index. To do so may require significant changes to the methodology. We also expect that as users begin to analyse the data, further questions that we may wish to add to WES may be developed.
Thus, we see WES very much as an evolving survey, whose methodology will have to adapt to changing requirements and to the knowledge we gain through experience with the survey. Maintaing this flexibility while maintaining the stability required by an ongoing longitudinal survey will be an interesting challenge.

REFERENCES

Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D qualifying paper. Harvard University.

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

Stahel, W.A. (1981). Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators. Ph.D. thesis (in German), ETH, Zurich.

RESUMÉ

Cet article décrit la méthodologie de l'Enquête sur le milieu de travail et les employés, une nouvelle enquête longitudinale de Statistique Canada qui collecte des données sur les lieux de travail, et ainsi que leurs employés.