

# Survey of Employment, Payroll and Hours: An Update

Chantal Grondin, Pierre Lavallée  
Business Survey Methods Division, Statistics Canada

## 1. Introduction

The Survey of Employment, Payroll and Hours (SEPH) is designed to provide monthly estimates to measure levels and month-to-month trends of payroll, employment, paid hours and earnings. The data are compiled at detailed industrial levels for Canada, the provinces and the territories. The target population is composed of all employers in Canada, except those primarily involved in agriculture, fishing and trapping, private household services, religious organizations and military personnel of defense services. This paper presents an update on the latest redesign that took place to convert SEPH to the new North American Industrial Classification System (NAICS).

## 2. Background

Since 1998, considerable savings are being realized with this redesigned survey. The number of employees and gross monthly payroll variables are now directly obtained each month from the complete file of payroll deductions remittance forms (PD7) from the Canada Customs and Revenue Agency (CCRA). To complement the data of this administrative source, a sample of approximately 10,000 establishments is used every month. This sample is used to collect information on employment, gross monthly payrolls, total hours, summarized earnings, as well as the breakdown of these variables by categories of employees (hourly paid, salaried and other). This sample is known as the Business Payroll Survey (BPS). The SEPH final estimates use a combination of both data sources. Regression models are used to predict hours and summarized earnings from the sample of BPS respondents. The resulting estimated regression coefficients are then applied to each record on the administrative source to mass impute hours and summarized earnings. Other variables (such as hourly employees or salaried employees) are obtained by multiplying employment, hours or summarized earnings by a ratio (or a function of ratios) estimated from the BPS sample. More details about the methodology of SEPH are available in Rancourt and Hidioglou (1998).

Starting in January 2001, the SEPH estimates are based on the NAICS. The migration to the new classification enhances the analytical possibilities of the SEPH data series since many other surveys in Statistics Canada are now NAICS based. In addition, the new classification is becoming the standard used by Canada's partners in the North American Free Trade Association (NAFTA), the United States of America and Mexico. The monthly historical SEPH series (from January 1991 to December 2000) were converted to the NAICS. At the same time, the revised series were adjusted to incorporate the new levels of employment and earnings derived from the inclusion of the large businesses in the administrative sample (phase III of the redesign in May 1998)<sup>1</sup>.

## 3. Current design

This section gives details on the current survey design. It is divided into three subsections, the first dealing with the administrative source, the second with the survey source and the third, with estimation using both sources.

---

<sup>1</sup> Since May 1998 and until December 2000, employment and payrolls growth rates were calculated from the administrative file and applied to the levels published in April 1998. This was to compensate the change of levels in the estimates that occurred when SEPH moved from phase II to phase III of its redesign.

### 3.1 Administrative files

#### Description:

In Canada, enterprises are required to remit to the Canada Customs and Revenue Agency (CCRA), deduction amounts retained from the employees' wages and salaries. Enterprises send their remittances for each list of employees using payroll deductions accounts (or *Business Number accounts* (BN)). Since 1993, an agreement exists between CCRA and Statistics Canada by which CCRA also has to collect from the enterprises their number of employees and total amount of payrolls (besides the remittances). Consequently, Statistics Canada has access to a complete monthly file containing remittances, number of employees and payrolls for all BN accounts. The number of in-scope BN records is approximately 1,000,000.

BN accounts are of two types: automatic (A) and twice monthly (TM). The "A" accounts represent remitters whose average monthly remittances in the previous year were less than \$15,000. Those whose remittances were less than \$12,000 are eligible to become quarterly remitters if this is what they want. All other A accounts remit once a month. On the other hand, the TM accounts represent remitters whose average monthly remittances in the previous year were \$15,000 or more. The TMs remit one to five times a month depending on the type of payrolls used (i.e., monthly, weekly, biweekly, etc.). For TMs, aggregation must be done to obtain a monthly figure for payroll, and to obtain the number of employees during the reference week (last seven days of the month). The aggregation process is often straightforward for the payroll variable. But it can be really complex for the number of employees, because one has to determine if employees are repeated from one form to the next and if so, make sure that they are not double-counted.

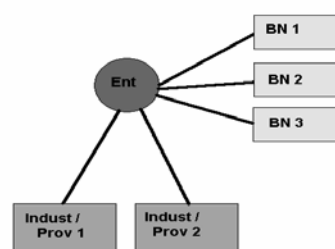
#### Edit and imputation:

BN account data undergoes a lot of editing. Large differences in reported employment and payroll between two months are identified and corrected within each industry (2-digit NAICS level). This process is referred to as the "Big Changers". Other outliers are also identified within strata (based on industry, groups of provinces and size). Large values of employment and payroll are identified using the quartile method while trend outliers and ratio outliers are identified using the Hidroglou-Berthelot method (1986). Trend and ratio outliers have their values of employment and payrolls set to missing so they can be imputed. Outliers with large values of employment or payroll are simply kept out of the pool of donors used to calculate trends and ratios for imputation.

Imputation is performed on the administrative data source, either for the number of employees, the payroll or both. Historical data is available for all records except births because a census of the administrative file is used each month. This historical data is used as often as possible to impute using the trend method. This method calculates trends in imputation classes and applies them to the previous month value to impute a current month value. If the previous month value is judged of poor quality or unreliable but that the information on the other variable (payroll if employment is being imputed, or vice versa), then ratio imputation can be used. Again, ratios are calculated within imputation classes. As a last resort, mean imputation is used. Imputation classes are comparable to strata, but are often at an even more detailed level of industry or province. In fact, in contrast to strata which are fixed in time, imputation classes can be changed. They are set up so that imputation is done at the most detailed level possible providing the number of observations needed to calculate the trends or ratios is high enough. If this is not the case, collapsing of imputation classes occurs until the minimum number of observations for imputing the data is attained. This number of observations has been set to five, based on an evaluation study.

#### Conversion of BN level data to establishment level data:

The information present on each BN record regarding industry and at the right level, especially if the BN belongs to an enterprise industry or province (multi enterprises). For multi enterprises, th



industry or province code to BN accounts as such, because BN accounts refer to groups of employees that could work in many different industries or provinces (as shown on the figure ). To solve this problem, a dominant establishment is identified in each multi enterprise based on the number of employees. The industry and province codes found on the BR for this dominant establishment is then assigned to all BN records associated to this enterprise. This information is the one used for outlier detection and imputation, but not for estimation.

For estimation, the information is assigned to the right industry and province. This is done in two steps. First, after imputation has been performed, the number of employees and payroll of the BN records is aggregated to the enterprise level. The second step consists in breaking down the enterprise information (employment and payroll) to the establishment level. This is straightforward for single enterprises (i.e., with only one establishment). For enterprises with more than one establishment, this is done using allocation ratios. These ratios are calculated based on the number of employees of each establishment from the BR, divided by the total number of employees for all establishments of an enterprise<sup>2</sup>. A record is then created for each establishment of an enterprise, containing industry and province codes, as well as the allocated number of employees and corresponding payroll. The resulting file is then basically ready for estimation.

#### Estimation:

Estimation of the number of employees and payroll is direct and uses the census of administrative records. Let  $U_A = \{1, \dots, k, \dots, N\}$  be the population of administrative data. The total  $X_A$  of administrative data for a variable of interest  $x$  (e.g. number of employees or total payroll) is obtained using  $X_A = \sum_{U_A} x_k$ .

### **3.2 The Survey source (BPS)**

#### Frame:

All establishments of the enterprises present on Statistics Canada's Business Register (BR) constitute the frame for the survey portion of SEPH. To be in-scope, an establishment must be an employer (have at least one employee), be active, and belong to any industry, except those primarily involved in agriculture, fishing and trapping, private household services, religious organizations and military personnel of defense services. Every month, a new frame is available: births are added, deaths are removed, and the design information is updated for all units, except for the in-sample take-some units. Every month, the frame contains approximately 900,000 establishments.

#### Model groups:

Model groups are used at the estimation stage. They form the level at which regression and ratio estimation are performed. Model groups are created by dividing the population of establishments into sub-populations within which efficient regression fits can be achieved. Model groups can also be seen as groups of strata. All units in the administrative and the establishment portions are assigned to the model groups. The sample of establishments is partitioned into  $G_E$  groups while the administrative records are partitioned into  $G_A$  groups, i.e.  $s_E = \bigcup_{g=1}^{G_E} s_{E,g}$  for the establishment sample and  $U_A = \bigcup_{g=1}^{G_A} U_{A,g}$  for the census of administrative records.

---

<sup>2</sup> Note that currently, these same allocation ratios are used to split the payroll even if in reality, these two variables do not follow the same proportions. This is causing important distortion in the estimates. However, work is underway to come up with a method to estimate different allocation ratios for payroll based on the BPS data.

This partition of both the establishment and administrative universes into regression groups is based on combinations of NAICS industries. Most model groups are at the 3-digit NAICS level, while most others are groups of 3-digit NAICS. In SEPH, there are 75 model groups. Since industrial activity is more discriminatory than geography, all model groups were kept at the national level with the current design, as opposed to some model groups being divided by region and sometimes size with the previous design. Note that estimates are only needed at the model group level because of the use of synthetic estimators.

Stratification:

Establishments whose number of employees (according to BR) is larger than a specified limit are called Take-All units (TA). These units are selected with certainty. They are further divided into two sets: the extreme ones (TA1) and the others (TA2). The TA1 units are excluded from the regular model groups and therefore excluded from the regression fits, as they could affect the regressions otherwise. Instead, each TA1 is assigned to its own unique model group. On the other hand, TA2 units are included in the model groups and the regression fits. The remaining units are assigned to take-some strata. Identification of take-all units is done using the *sigma-gap* method. This procedure is carried out as follows. Let  $x_{(k)}$  be the  $k^{\text{th}}$  ordered value for the variable “number of employees”. A unit  $k$  is declared a take-all unit with respect to variable  $x$  if

- 1)  $x_{(k)} > x_M$                       where  $x_M$  is the median; and
- 2)  $x_{(k)} - x_{(k-1)} > a\sigma_x$       where  $\sigma_x$  is the standard deviation and  $a$  is a specified constant.

All units  $k'$  such that  $x_{(k')} > x_{(k)}$  are also take-all units. The sigma-gap method used in SEPH is a slightly modified version of the above method in that the 5% largest records are not used to compute the sigma ( $\sigma$ ) value. In SEPH,  $a = 7$  for TA1 units and  $a = 0.75$  for TA2 units.

The model groups being the sole domains of interest for the BPS, stratification is done within each one of them. Stratification divides the model groups into size categories and region, and is done independently for each model group, yielding a different number of strata in each model group. The size categories are defined as: 1-19 employees, 20-49 employees, 50-99 employees and 100 or more employees. To ensure a minimum number of units in each stratum, size strata were sometimes collapsed. The regions are defined as: Atlantic provinces, Quebec, Ontario, and the rest of Canada including the Territories. Again, not all model groups were divided by region and collapsing was sometimes necessary.

Sample allocation and selection:

The objective of the sample design for the BPS is to produce reliable estimates of regression coefficients and ratios at the model group level. With this goal in mind, Neyman allocation was used to allocate the sample. This was done in two steps: first, we allocated the sample to each model group, and second, to each stratum within the model groups. Allocation to model groups took into account the variability of six different variables as opposed to only one variable with the previous design. These six variables were: average weekly earnings for salaried employees, average hourly earnings for salaried employees, average weekly earnings for hourly employees, average hourly earnings for hourly employees, number of hours excluding overtime for hourly employees and total number of hours including overtime for hourly employees. The final sample size for each model group was chosen to satisfy the needs of most of these six variables. Then, the sample was allocated to strata within model groups based on the ratio of the number of hourly employees to the total number of employees. The total sample size of the BPS is approximately 10,000 units, yielding coefficients of variation of approximately 15% for the most important variables of interest.

SEPH uses the Generalized Sampling System (GSAM) to select and rotate the BPS sample. One twelfth of the sample rotates in and out every month in the take-some strata. GSAM uses the collocated sampling method. For the NAICS redesign, model groups and strata were completely redefined based on the NAICS. To avoid unwanted fluctuations in the BPS estimates between the previous SEPH design and the new one (between December 2000 and January 2001), the NAICS sample was drawn with a maximum overlap with the SIC sample. The achieved overlap rate was 54%.

#### Exclusion process:

A new process has been put in place to try to increase the sample size without increasing collection costs. It is called the exclusion process. This process consists in identifying units in the sample for which we either know that we will not get any data (hard refusals), we know that the company is out of business (but has not yet been removed from the BR frame), or we know that there is a high chance that the company has no employees for this month. Once these units have been identified, they are removed from the files that are sent for data collection and treated separately. The hard refusals are coded as such, while the others are imputed to zero employees. Because these exclusions decreased the number of units sent for data collection and that we have the financial resources to collect approximately 10,000 units, sampling fractions were increased the following month based on the global exclusion rate. In the near future, exclusion will be calculated at the stratum level to be more optimal. So far, this process has allowed an increase of approximately 700 usable units.

#### Data collection:

Data collection is done by mail or fax in 70% of the cases, and by telephone in almost all other cases. A very few cases are Electronic Data Reporters (EDR) at the moment, and work is underway to attempt to increase that number in the near future. This would be another mean of increasing sample size without increasing collection costs. The software currently used for data collection is CASES, but it will be replaced by BLAISE next year. Note that questionnaires received by mail or fax are entered in CASES just as if they were telephone interviews. Edits that pop-up on the screen for these questionnaires are addressed by following up with the respondents over the telephone.

In April 2001, historical edits were implemented for the first time in the application. In the past, the BPS was always seen strictly as a cross-sectional survey. Since 11/12 of the units are the same between two months, and fluctuations are often seen between two months in the estimates, it was desired to obtain more consistency in the micro-data between two months. This was achieved with the historical edits.

#### Edit and imputation:

Once collection is over for a given month, a series of edits take place. The first series of edits identify the largest differences for each of the variables of a record from one month to the next. The records with the largest differences are identified and manually corrected. As well, outlier detection is performed in each stratum using the quartile method. Records identified as outliers have their final weight set to one. As well, regression outliers are identified using the Cook distance statistic. All records with a Cook distance greater than five are considered as regression outliers. These outliers have their regression weights set to a very small value so that their impact on the regressions is minimal. More details on weight calculation will follow.

Before the NAICS redesign, the only imputation done in the BPS was for TA1 take-all units. Since January 2001, what started as a quick-fix contingency plan to impute as many non-respondents as possible in the case of low response rates, has remained in place. Non-respondents in the current month that have information in one of the last three months (responded or imputed) are now imputed using the latest information. No adjustment is done to the imputed data at the moment.

However, special care is taken to avoid using information related to the month of December since this month is known to include many special payments. Records that cannot be imputed are simply treated as non-respondents and a weight adjustment is done to account for them. More work should be done in the coming years to improve imputation in the BPS.

#### Weighting:

There are two series of weights used in the BPS: sampling weights used to estimate ratios (e.g. proportion of hourly employees, proportion of earnings coming from overtime) and regression weights used in the modeling of hours and earnings from employment and payroll. For stratum  $h$ , the design weights for the establishment sample ( $E$ ) is calculated as:

$$w_{Eh} = (N_h / n_h) \times (n_h / n_{h,resp})$$

The regression weights are calculated as:

$$v_{E,hours,i} = w_{Eh} / emp_i \quad \text{for the model for hours and}$$

$$v_{E,summ,i} = w_{Eh} / pay_i \quad \text{for the model for summarized earnings.}$$

Note that the regression weights are equal to the design weights divided either by the number of employees or the total payroll of the record (depending on the model) to get rid of the effect of heteroscedasticity of the residuals.

### **3.3 Estimation using both sources**

The establishment sample uses a model-assisted approach for the estimation of totals and ratios for variables not collected by the administrative portion. Models involving linear regression are estimated within each model group. The number of employees and total payrolls for the month are the independent variables, while total hours and summarized earnings are the dependent variables. Ratios of employment, total hours and earnings by category of employees are also estimated within each model group. The number of employees and total payrolls reported on the administrative portion are used to construct totals of auxiliary variables. Estimated parameter values from the regression are used to predict total hours and summary earnings for each administrative unit in the model group. That is for each unit, the missing variable is imputed using mass imputation, as described for example in Kovar and Whitridge (1995).

The estimation process may be viewed as a two-step regression procedure. In the first step, the auxiliary population totals are derived. For a given domain  $d$ , the totals (for employment and payrolls) are obtained directly from the administrative portion  $U_A$  as

$$X_A(d) = \sum_{U_A} x_{A,k}(d)$$

In the second step, the estimated regression coefficients are obtained from the fit between the available variables on the BPS sample, for each model group. For the regression of the dependent variable  $y_k$  and the vector of independent variables  $\mathbf{x}_{E,k}$ , we have

$$\hat{B}_{E,g} = \left( \sum_{s_{E,g}} v_{E,k} \mathbf{x}_{E,k} \mathbf{x}'_{E,k} / \hat{\sigma}_{E,k}^2 \right)^{-1} \sum_{s_{E,g}} v_{E,k} \mathbf{x}_{E,k} y_k / \hat{\sigma}_{E,k}^2$$

The values  $\mathbf{x}_{E,k}$  are the data corresponding to the  $\mathbf{x}_{A,k}$  data from the administrative source, and  $\hat{\sigma}_{E,k}^2$  is a variance factor. The dependent variables in these fits are either hours or summarized earnings. The maximum number of independent variables included in the fit is two: payroll and/or number of employees. No intercept term was included as the presence of an intercept was found to increase the possibility of obtaining negative regression weights. Finally, the predicted variables are

produced using: 
$$\hat{Y}(d) = \sum_{g=1}^G \sum_{U_{A,g}} \mathbf{x}'_{A,k} \hat{B}_{E,g}$$

#### Variance estimation:

Variance calculation has improved with the NAICS design. It now takes into account the variability from both sources of data: the sampling variability from the BPS (with a Jackknife) and the imputation variance for the administrative source (model-assisted approach). Work is underway to incorporate the variance due to the regression models that are used to mass impute hours and summarised earnings.

#### Confidentiality:

The disclosure risk method used in SEPH is data suppression. An annual confidentiality pattern is created based on the previous year's twelve months of microdata. The CONFID software is used to generate a confidentiality pattern for each of the twelve months. Then the twelve patterns are compared and a cell is declared publishable in the annual pattern only if it was judged publishable for each of the twelve months. The use of twelve months of microdata takes into account the seasonality of the industries and allows the users to have access to a series of data for at least the whole year (since the same pattern is used for the complete calendar year).

#### **4. Conclusion**

The SEPH design has changed significantly during the last seven years. It went from a survey-based design with a monthly sample of 70,000 establishments, to a design combining administrative data and survey data with a monthly sample of approximately 10,000 establishments. Consequently, collection costs were substantially reduced. Many improvements were brought to the survey processes with the NAICS design. However, more improvement is still required in several of them. The next few years will be devoted to implementing them and monitoring the survey processes so as to keep the estimates at their highest quality level possible.

#### **5. References:**

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

Kovar, J.G., and Whitridge, P.J. (1995). Imputation of business survey data. In *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. eds, 403-423, J. Wiley and Sons.

Rancourt, E. and Hidiroglou M. (1998). *Use of Administrative Records in the Canadian Survey of Employment, Payrolls, and Hours*, Proceedings of The Survey Section of the Statistical Society of Canada, 39-49.