Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

The Challenges of the Use of Administrative Data in the Survey of Employment, Payrolls and Hours

by Anthony Yeung, Anne-Marie Houle, and Sharon Wirth

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

The Challenges of the Use of Administrative Data in the Survey of Employment, Payrolls and Hours

Anthony Yeung, Anne-Marie Houle, and Sharon Wirth[1]

Abstract

The Survey of Employment, Payrolls and Hours (SEPH) is a monthly survey using two sources of data: a census of payroll deduction (PD7) forms (administrative data) and a survey of business establishments. This paper focuses on the processing of the administrative data, from the weekly receipt of data from the Canada Revenue Agency to the production of monthly estimates produced by SEPH.

The edit and imputation methods used to process the administrative data have been revised in the last several years. The goals of this redesign were primarily to improve the data quality and to increase the consistency with another administrative data source (T4) which is a benchmark measure for Statistics Canada's System of National Accounts people. An additional goal was to ensure that the new process would be easier to understand and to modify, if needed. As a result, a new processing module was developed to edit and impute PD7 forms before their data is aggregated to the monthly level.

This paper presents an overview of both the current and new processes, including a description of challenges that we faced during development. Improved quality is demonstrated both conceptually (by presenting examples of PD7 forms and their treatment under the old and new systems) and quantitatively (by comparison to T4 data).

Key Words: Administrative data, Edit, Imputation.

1. Introduction

The Survey of Employment, Payrolls and Hours (SEPH) is designed to provide monthly estimates of levels and month-to-month trends of employment, payroll, paid hours and earnings at industrial levels for Canada, the provinces and the territories. The industrial levels considered are based on the North American Industrial Classification System (NAICS). The target population is composed of all employers in Canada who make deductions from employees' wages and salaries to cover income tax, Canada Pension Plan contributions and Employment Insurance premiums. All industrial sectors are included except those primarily involved in agriculture, fishing and trapping, private household services, religious organizations and military personnel of defense services.

The data for SEPH is collected using two different data sources: a census of administrative data for the employment and total payroll variables and the data obtained from a monthly survey. Some variables are obtained directly from the administrative source while others are estimated using both sources.

As is the case for all surveys, validation and imputation is required to ensure that the final data used for calculating the estimates is complete, consistent and valid. This is particularly relevant for the administrative data source used by SEPH because this data is collected for non-statistical purposes. The current processes were analyzed and a new approach was proposed to identify and correct errors on the PD7 forms before aggregating to monthly values.

[1]Anthony Yeung, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, K1A 0T6, anthony.yeung@statcan.gc.ca; Anne-Marie Houle, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, K1A 0T6, anne-marie.houle@statcan.gc.ca ; Sharon Wirth, Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, K1A 0T6, sharon.wirth@statcan.gc.ca

This paper focuses on the receipt and processing of the administrative (PD7) data. Section 2 contains descriptions of the information collected on the PD7 forms, of the current data processing and of the reasons behind the decision to modify the process. Section 3 focuses on the challenges of developing the new edit processing system. Section 4 gives a detailed description of the new process and some results. Finally, a summary of the achievements and the next steps in development are given in Section 5.

2. Processing of the administrative data

2.1 SEPH's data sources

The first data source, payroll deductions remittance forms (PD7), is a census of administrative data received from the Canada Revenue Agency (CRA). There are two variables of interest on these PD7 forms: the number of paid employees in the last pay period of the month (E_F), and the gross amount paid to employees in the remittance period (P_F). A third variable, the amount remitted to CRA by the employer for Income Tax, Canada Pension Plan and Employment Insurance on the employee payroll (R_F), is also available from the PD7 forms and is highly correlated with P_F.

The other data source is a monthly establishment survey, the Business Payrolls Survey (BPS). A sample of about 11,000 establishments selected from the Statistics Canada's Business Register is used to collect information on employment, payroll, total regular and overtime hours, special payments, as well as a breakdown of these variables by categories of employees (paid hourly, salaried and other).

The SEPH final estimates use a combination of both data sources. The estimates of the gross monthly payroll and of the number of employees are obtained directly from the administrative data. Regression models are used to predict hours and weekly 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 weekly earnings. Other variables, such as hourly employees, are obtained by multiplying employment, hours or weekly earnings by a ratio estimated using the BPS sample. More details about the methodology of SEPH are available in Rancourt and Hidiroglou (1998).

2.2 Description of the administrative data source

As mentioned in Section 2.1, three variables are obtained from the PD7 forms: the number of paid employees in the last pay period (E_F), the amount paid to employees in the remittance period (P_F) and the amount remitted to CRA by the employer (R_F). A set of complex aggregation rules is used to obtain the monthly values that are required by SEPH. The complexity is a result of the various remittance patterns.

Each PD7 form represents a single remittance of funds and is associated with only one enterprise. However, one enterprise may submit separate PD7 forms for different sets of employees (including remittance periods of various lengths), depending on its payroll practices and history. According to CRA rules, employers with historically large annual remittances are required to remit funds a minimum of two times per month. Medium-sized employers are required to remit once per month, while those with the smallest remittances are allowed to remit quarterly.

2.3 Current processing of the administrative data after aggregation

In the current production processing system, only a few simple data validations are performed at the form level, such as the identification of forms with unusual remittance to pay ratio. Most of the processing of the administrative data is done after aggregation to the monthly level. The current processing system includes automatic corrections, outlier detection, manual corrections and imputation.

Examples of edit rules applied at the monthly level are the identification of implausible P_A/E_A ratios[2] according to industry-specific limits and the identification of extreme month-to-month trends of P_A or E_A.

Outlier detection is also performed within each stratum (strata are based on industry, groups of provinces and employment size). The quartile method is used to flag large values of employment and payroll; flagged records are excluded from the calculation of trends and ratios used for imputation. The Hidiroglou-Berthelot (1986) method is used to identify outliers in trends and ratios. If records are identified as outliers for the P_A/E_A ratio, both P_A and E_A are then set to missing and flagged for imputation; the other ratio outliers and the trend outliers are only flagged for exclusions. Finally, the enterprises with the largest differences in reported level of payroll and/or employment between the current month and the previous month are identified within each industry. Analysts validate these records and manually edit them if required.

Imputation is performed either for E_A, P_A or both. A few methods are used depending on the information available. If the unit appears to be active in the current month and if the previous month data is of good quality then trend imputation is used. If usable data is not available for the previous month but other current month variables are available (P_A if E_A is being imputed or vice versa), then ratio imputation is used. As a last resort, current month stratum averages are calculated and used for mean imputation.

2.4 Reasons for modifying the administrative data processing

There were indications from the System of National Accounts (SNA) of inconsistencies between the payroll annual growth rates obtained from the T4 data and from the aggregated annual PD7 data. T4 data contains the total annual income paid to the employees for each employer. To investigate that situation, the payroll level and trends were produced based on the raw data received from CRA and on the data after processing. This comparison indicated a reduction of payroll level and a smaller month-to-month increase after processing. An evaluation of the SEPH administrative data processing was therefore undertaken to identify which of the processing step(s) resulted in these reductions.

This evaluation led to a suggestion by the survey analysts that some editing at the form level, before aggregation to the monthly level, may be beneficial. Indeed, there is a strong correlation between the gross amount paid (P_F) and the remittances (R_F), as well as a known relationship between these variables due to income tax laws. Moreover, some obvious reporting errors are more easily (or only) identifiable and explainable at the form level.

An ulterior goal was to standardize and automate the editing and imputation process as much as possible, as there is a large volume of data to process every month. The existing process consists of many steps, is very complex and is not easy to modify. Modifying the system provided an opportunity to improve in these areas.

The administrative data processing was therefore reviewed completely. As a result, a new module for edit and imputation at the form level was developed. This new module is mainly for editing and imputing the payroll variable. As is explained in Section 4, most of the processing of the employment variable is done at the aggregated level.

3. Challenges in the development of the form-level edit and imputation module

Several challenges were faced while developing the new form-level Edit and Imputation (E&I) system. These include the different reporting patterns, the various types of errors that can occur during data collection, as well as the use of the annual T4 income as a benchmark to evaluate the new process.

[2] The variable notation used is similar to that described in section 2.1. The subscript A denotes the values after the aggregation rules have been applied, whereas the subscript F denotes form-level values.

3.1 Identification of reporting patterns

In order to develop a form-level E&I process and to identify which records demonstrate irregularity in the data, it was necessary to understand the relationship between the variables on the PD7 forms and other administrative sources. With around one million PD7 forms received per month from CRA, it was a big task to identify many reporting and errors patterns. Since there is no standard on the number of forms and the coverage of each form for the reporting of money to CRA (only on the frequency of remittance), this leads to many possible ways of remitting. For example, an employer who is required to remit once a month can choose to send PD7 forms to CRA each time their employees are paid (twice a month, bi-weekly, weekly, etc.). In addition, if the employer has different categories of employees such as full-time and part-time employees or salaried and hourly employees, the employer might fill out separate forms for each category. Based on the income tax rules, one expects that the remittance rate for some categories of employees will be much lower than for the other categories within the same enterprise.

3.2 Types of errors in the data collection process

Since the data collected is a census of the PD7 forms, the problems identified on each form are due to non-sampling errors from different sources. The following types of errors were identified during the development of the form-level E&I process.

a. Transcription errors

These types of errors occur when the employers fill out the PD7 forms. Some employers might enter the information incorrectly (e.g. pay and remittance fields are switched, cents are included in the pay and/or remittances, months and days are switched, etc.). These types of errors can be identified through the proposed form-level edit system.

b. Conceptual misinterpretation of the variables on the PD7 form

The guideline for filling out the PD7 form is not clearly indicated on the form. For the value of pay, the employers are to include all remuneration that they pay to the employees, including taxable benefits and allowances as well as special payments, but these payments are sometimes not included on the PD7 forms. These types of errors cannot be caught by the proposed form-level edit system because the system is designed to identify obvious errors. These problems might only be identified using other administrative sources such as the T4. Another example is that for the quarterly remitters, some employers might include the entire amount paid to employees during the entire 3-month period even though the amount requested is for the last month pay.

c. Data capture error at CRA

All the forms sent to CRA are being captured at CRA and errors can occur during the data capture process. Occasionally errors in remittances may occur, but an internal accounting check is done by CRA to compare the remittances reported on the form to the actual dollar amounts remitted by the employers. However, no validations are done for employment and payroll at CRA. Consequently, errors in remittances are more likely to be identified and corrected by CRA than are errors in the employment and pay fields of the PD7 forms. For these two fields, some obvious mistakes can be identified through the proposed form-level edit system.

d. Errors introduced by the "middleman"

Some employers hire payroll service companies or use financial software to complete the remittance forms. It was discovered that in some occasions, systematic data errors occur because of programming or transcription errors. Some of these errors are identified by the current quality assurance procedures and actions are taken to correct the mistakes.

e. Data transmission errors (duplicate transmissions due to unknown reasons)

Duplicate forms are sometimes erroneously transmitted to Statistics Canada. The proposed edit system can identify these duplicate forms.

f. Not all variables are reported

Although the PD7 form requests the information on the pay and the number of employees, a number of employers are only indicating the remittance amounts on the form. Because of this, an imputation method has to be applied to provide estimates of the variables missing on the form. With the proposed form-level E&I system, the gross pay is imputed for every form and the number of employees is imputed at the aggregated monthly level.

3.3 Use of the T4 data as benchmark

The increase in consistency between the aggregated annual PD7 and the T4 data is one of the main objectives of the redesign of the administrative data processing. Moreover, the T4 data is the only source of comparison that can be used to assess the gain of using a form-level E&I module. While it is widely believed that these two sources provide an exact or a very close comparison of the annual income, it is not entirely correct.

A study, performed to measure the expected discrepancy between the two sources, showed that the annual PD7 pay is between 1.4% and 1.8% lower than the T4 annual pay for the years 2003 to 2005. The study was based on the records that were not modified at all by the administrative data processing. In other words, these records had non-missing data that satisfied all of the data assumptions at the form level.

This study also helped to explain the differences between the two sources. First of all, the aggregated annual PD7 pay may not include data for all 12 months in a given year, while T4 pay is a true annual total. For example, a monthly PD7 value may have been dropped from the SEPH file because the unit was not in-scope for SEPH for a given month. Secondly, for records that are in the two sources' universes, there are also various explanations why the T4 total is greater than the aggregated annual PD7 total. The reasons identified are listed below:

a. Some special payments included in the T4 pay may not be included in the gross monthly payroll reported on the PD7 form. For example, records were found where a very large remittance value is reported on the PD7 forms in a specific month every year while the reported pay is only the regular pay;

b. Even though some taxable benefits, such as the personal use of the employer's automobile, should be included in the gross payroll amount reported on the PD7 forms, some companies may not include them on the forms. Examples were found where the PD7 pay is smaller than the T4 pay because of this situation;

c. Employers have until the end of February of Year Y+1 to report remittances to CRA for Year Y. If some remittance amounts for Year Y are reported in the first two months of Year Y+1, this can result in an underestimation of Pay for Year Y on the PD7 forms;

d. If enterprises do not receive their remittance form from CRA for a month, if they lose it or if they discover they made an error in remitting their source deductions, they are asked to send a cheque or money order to their tax centre with a note indicating the pay period to which the amount applies. Any amounts remitted to CRA without a PD7 form contribute to the underestimation of the PD7 pay compared to the T4 pay.

It was not always possible to identify the records in these situations but some were identified. If these records are excluded from the analysis, the difference between the two sources is reduced significantly and even eliminated for some years. So, although it is believed the two data sources are identical, there are some non negligible differences between them. These differences must be taken into consideration when evaluating the new E&I module.

4. Form-level edit and imputation process

4.1 Methodology

On the PD7 form, remittance is more accurate than either employment or pay because some validation on remittances is done by CRA as described in section 3.2. Because of the high correlation between remittance (R_F) and pay (P_F) and because of the known tax relationship between these variables, editing pay at the form level could be highly successful.

In contrast, it is difficult to edit employment at the form level. For any given enterprise, the total number of employees is usually fairly stable from one reference month to another and errors in employment are often easy to identify after aggregation to the monthly level. However, this is not always the case from one PD7 form to another for a given enterprise. This is because an enterprise may remit separate forms for different categories of employees. Since the relationship between E_F and P_F or between E_F and R_F may be very different from one category to another, this makes it difficult to identify errors in E_F at the form level. Consequently the focus of the form-level edit rules is the editing of pay but not all problems can be identified at that level. An edit and imputation process is still required at the aggregated monthly level, especially to identify potential problem with employment.

4.1.1 Data assumptions

The system developed is based on a set of explicitly-stated assumptions about the content and the relationship between variables P_F, E_F and R_F. For example, we assume that E_F is an integer and $E_F > 0$, and $P_F > E_F$. Considering that the annual remittance due is defined by federal and provincial tax laws, additional data assumptions about the ratios R_F/P_F and P_F/E_F can be defined. For example, for the province of Quebec, the following assumption is used: $0.05 < R_F/P_F < 0.4$. For provinces outside of Quebec, the following assumption is used: $0.1 < R_F/P_F < 0.6$.

Each form is evaluated to determine if it passes or fails each assumption. The population of forms is then divided up into mutually exclusive and exhaustive groups. For each group, the minimum change required to change the least number of variables in order to satisfy the assumptions is determined and a set of edit rules are applied.

4.1.2 Form-level edit rules

There are two categories of edit rules: the validity edits and the consistency edits. Validity edits deal with only one variable on a form and are used to identify situations where the variable is outside an acceptable range. For example, because remittance on the form is always available and positive, the pay field has to be greater than 0. Consequently, if pay on the form is 0, it is set to missing and will be imputed later. Consistency edits deal with relationships between variables on a form. Some of the consistency edits only use the information of the form being edited while other edits use historical data (i.e., the information from previous forms for the same group of employees). An example of a consistency edit not using historical information is to set P_F to missing and flag it for imputation if $P_F < R_F$.

The validity edits are applied first followed by the consistency edits using historical information and then by the consistency edits using information on the form. The order of priority for applying the historical consistency edits is based on the concept of a distance measure. This distance measure is defined in terms of the perceived reliability of the three variables and in terms of the number of variables used for matching. Historical edit rules that use three variables are applied first. Edit rules that use two variables are applied second. Since R_F is considered to be more reliable than P_F, edit rules based on matching to R_F and E_F have a higher order of priority for application than rules based on matching to P_F and E_F. Edit rules that match to one variable are applied next. It was felt that R_F was the only variable reliable to use in matching criteria with one variable only. Finally, all forms with obvious problems in all three variables P_F, E_F and R_F are automatically sent for manual processing.

Some research was done in order to determine the best choice for the maximum number of forms to use in attempted matching as well as the maximum time lapse between matching forms. It was decided to allow matching attempts to a maximum of 48 forms with a maximum elapsed time of one year.

More information about the methodology of the form-level edit process can be found in Wirth (2007).

4.1.3 Form-level imputation process

Form-level imputation is used to find the best estimate of P_F when pay cannot be determined during the edit process. Only P_F is imputed for each form whereas the number of employees is imputed at the aggregated monthly level if required.

Seven imputation methods for P_F are suggested at the form level. As is the case for the editing process, the priority of the methods being applied is based on the concept of a distance (reliability) measure. The methods utilizing more accurate historical information are applied first since it is believed that they produce a better estimate for pay. For example, if a close match on the employment and remittance values is found based on the historical forms that did not fail any edit, the information (E_F, P_F, R_F) of the matched form is used to impute the pay of the current form (Method 1). If there is no close match based on the employment and remittance values, the next available imputation method is based on a close match on remittances only (Method 2). If no close match is found, the next three imputation methods use ratios based on aggregated historical information of pay, expected pay or remittances (Methods 3, 4 and 5) to impute pay for the current form where the expected pay for a given employment and remittance level on each form is derived based on the provincial tax model. If a business has no history of reporting pay on its forms, pay is then imputed using the expected pay (Method 6). The final imputation method is used only when there is no history of reporting and no employment on the current form. With this method, P_F is imputed using the remittance value divided by one of two remittance-to-pay ratios, depending on the province of employment (Method 7).

It is noteworthy that after P_F is imputed, post-imputation edits ensure that the imputed pay satisfies the data assumptions presented in Section 4.1.1.

4.2 Comparison to the T4 data

As already mentioned, one of the main purposes of the addition of a form-level processing module is to increase data quality as well as consistency with the T4 data. In order to evaluate if this goal is reached, a comparison of the annual pay from the processed PD7 file and from the T4 file was performed at the Canada/all-industry level. This was done by computing the percentage difference between the total annual pay from the two sources for the records common to both sources. This percentage is calculated as $(P_{PD7} - P_{T4})$ x 100 / P_{T4} , where P_{PD7} and P_{T4} are respectively the annual pay from the PD7 file and from the T4 file.

The comparison was done for pay at two stages of the process: (i) after aggregation to the monthly level (in other words, after the form-level processing but before the monthly level processing) and (ii) after the monthly level processing.

In the current system, only a few automatic edits are performed at the form level. Most of the Edit and Imputation process is done at the monthly level, including manual corrections.

With the new system, many edits are added at the form level and the imputation of pay is performed at that level if possible. Imputation at the monthly level is then performed for the companies with no form received for the entire month. There is neither automatic nor manual correction performed at that level.

The new system is considered to be producing better results than the current system if the percentage difference between the aggregated annual PD7 and the T4 annual pay is smaller with the new system than with the current one. Tables 4.2-1 and 4.2-2 present respectively the percentage differences with the current processing system and with the new system.

Table 4.2-1
Percentage difference between PD7 pay and T4 pay – Current administrative data processing

Stage in the data processing	Year		
	2003	2004	2005
	Percentage difference		
Pay after aggregation	-10.3%	-10.2%	-8.1%
Pay after monthly level processing	-9.3%	-9.5%	-10.1%

Table 4.2-2
Percentage difference between PD7 pay and T4 pay – New administrative data processing including form-level E&I module

Stage in the data processing	Year		
	2003	2004	2005
	Percentage difference		
Pay after aggregation	-4.7%	-5.1%	-3.8%
Pay after monthly level processing	-2.0%	-1.7%	-1.1%

It is observed that the percentage differences are much smaller with the new form-level module applied (Table 4.2-2) than with the current system (Table 4.2-1). Moreover, the results mentioned in Section 3.3 indicate that for the records not modified by the new administrative data processing, the aggregated annual PD7 annual pay is lower than the T4 annual pay by about 1.8% in 2003 and 1.4% in 2004 and 2005. Since the difference for the unmodified records is very similar to those of the last line of Table 4.2-2, it is believed that the new form-level processing module greatly increases the coherence between the data sources.

5. Summary

A form-level edit module was developed to complement the current monthly level process to identify and correct any obvious and explainable reporting errors at the earliest stage possible. In addition, a form-level imputation process provides a realistic value of pay for each PD7 form based on many factors such as the historical reporting pattern of the enterprise and income tax regulations. This enhancement greatly improves the coherence of the monthly pay estimates when comparing with other data sources. Additional studies will need to be completed to further improve the data quality of the number of employees at the monthly level. Unfortunately there is no T4 like source to act as benchmark in this case.

The development of this new process sheds light on the remitting patterns of the employers. Statistics Canada now has the permission to contact the employers directly to get clarification on the data provided. This knowledge will help further improve the edit specifications.

References

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

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

Wirth, S. (2007). SEPH PD7 Edit System: Explanation of the Editing Concepts, unpublished report, Ottawa, Canada: Statistics Canada.